Global evaluation of runoff from ten state-of-the-art hydrological models

Hylke E. Beck[1], Albert I.J.M. van Dijk[2], Ad de Roo[3], Emanuel Dutra[4,5], Gabriel Fink[6], Rene Orth[7], and Jaap Schellekens[8]

[1]Princeton University, Civil and Environmental Engineering, Princeton, NJ, United States
[2]Fenner School of Environment & Society, Australian National University (ANU), Canberra, Australia
[3]European Commission, Joint Research Centre (JRC), Via Enrico Fermi 2749, 21027 Ispra (VA), Italy
[4]European Centre for Medium-Range Weather Forecasts (ECMWF), Redding, UK
[5]Instituto Dom Luiz, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisbon, Portugal
[6]Center for Environmental Systems Research (CESR), University of Kassel, Kassel, Germany
[7]Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland
[8]Inland Water Systems Unit, Deltares, Delft, The Netherlands
Correspondence to: Hylke E. Beck (hylke.beck@gmail.com)

Abstract. Observed streamflow data from 966 medium sized catchments (1000–5000 km^2) around the globe were used to comprehensively evaluate the daily runoff estimates (1979–2012) of six global hydrological models (GHMs) and four land surface models (LSMs) produced as part of Tier-1 of the eartH2Observe project. The models were all driven by the WATCH Forcing Data ERA-Interim (WFDEI) meteorological dataset, but used different datasets for non-meteorologic inputs and were run at various spatial and temporal resolutions, although all data were re-sampled to a common $0.5°$ spatial and daily temporal resolution. For the evaluation, we used a broad range of performance metrics related to important aspects of the hydrograph. We found pronounced inter-model performance differences, underscoring the importance of hydrological model uncertainty in addition to climate input uncertainty, for example in studies assessing the hydrological impacts of climate change. The uncalibrated GHMs were found to perform, on average, better than the uncalibrated LSMs in snow-dominated regions, while the ensemble mean was found to perform only slightly worse than the best (calibrated) model. The inclusion of less accurate models did not appreciably degrade the ensemble performance. Overall, we argue that more effort should be devoted on calibrating and regionalizing the parameters of macro-scale models. We further found that, despite adjustments using gauge observations, the WFDEI precipitation data still contain substantial biases that propagate into the simulated runoff. The early bias in the spring snowmelt peak exhibited by most models is probably primarily due to the widespread precipitation underestimation at high northern latitudes.

1 Introduction

Hydrological models are indispensable tools for many purposes, including but not limited to, (i) flood and drought forecasting, (ii) water resources assessments, (iii) assessing the hydrological impacts of human activities, and (iv) increasing our understanding of the hydrological cycle. It is more than 50 years since the first attempts at hydrological modeling (Lins-

ley and Crawford, 1960; Rockwood, 1964; Sugawara, 1967; Freeze and Harlan, 1969). Since then, a plethora of conceptual, physically-based, and stochastic hydrological models has been developed, each with its own assumptions and characteristics (for non-exhaustive overviews, see Singh, 1995; Singh and Frevert, 2002; Rosbjerg and Madsen, 2006; Trambauer et al., 2013; Sooda and Smakhtin, 2015; Bierkens et al., 2015; Kauffeldt et al., 2016). Because all hydrological models are inevitably imper-
5 fect representations of reality, they produce highly uncertain estimates even if we would have access to perfect meteorological data (Beven, 1989).

The quantification of these uncertainties using independent data sources is of critical importance to advance model development, reject deficient model structures and parameterizations, quantify model credibility, and ultimately bring some order in the plethora of models (Klemeš, 1986; Wagener, 2003; Döll et al., 2015; Clark et al., 2015). There have been several collabora-
10 tive research efforts focusing on the intercomparison and verification of hydrological models. The earliest were coordinated by the World Meteorological Organization (WMO, 1975, 1986, 1992). Other noteworthy initiatives include the Model Parameter Estimation Experiment (MOPEX; Duan et al., 2006), the Global Soil Wetness Project (GSWP; Dirmeyer, 2011), the Water Model Intercomparison Project (WaterMIP; Haddeland et al., 2011), and the Global Energy and Water Exchanges (GEWEX) LandFlux project (McCabe et al., 2016). These initiatives have led to numerous multi-model evaluation studies focusing on
15 such hydrological variables as runoff (e.g., Gudmundsson et al., 2012a; Zhou et al., 2012), evaporation (e.g., Schlosser and Gao, 2010; Jiménez et al., 2011; Miralles et al., 2015), soil moisture (e.g., Guo et al., 2007; Xia et al., 2014), snow cover (e.g., Slater et al., 2001), and total water storage (Güntner, 2008), among others.

One of the most useful variables for hydrological model evaluation is runoff, since it reflects the integrated response of a host of hydrological processes occurring in a catchment (Fekete et al., 2012) and because observations are readily available
20 for many catchments across the globe (Hannah et al., 2011). Table 1 lists, to our knowledge, all macro-scale (i.e., continental to global scale) studies evaluating the runoff estimates of multiple models that have been published so far. Out of these 20 studies, two focused on the conterminous USA, five focused on Europe, while thirteen had a global scope. However, many of these studies used observations from a relatively small number (< 100) of large catchments ($\gg 10\,000$ km^2). The use of a small number of basins limits confidence in the results and precludes a spatially detailed assessment, while the large size of
25 the catchments makes it more difficult to distinguish between deficiencies in the forcing, the (sub-)surface component, or the river routing component of the modeling chain. Moreover, a large number of the studies only evaluated monthly mean runoff, precluding analysis of the shape of individual flow events, or used the Nash and Sutcliffe (1970) efficiency (NSE), which has been criticized in several previous studies for being overly sensitive to the timing and magnitude of peak flows (Schaefli and Gupta, 2007; Jain and Sudheer, 2008). Furthermore, many studies considered only a few hydrological models (≤ 5) or
30 performance metrics (≤ 2), limiting the insights that can be gained.

As part of Tier-1 of the eartH2Observe project, ten state-of-the-art hydrological models were run globally at a daily time step for the period 1979–2012 using the same forcing dataset, in an effort to develop a global reanalysis of water resources that supports efficient water management and decision making (Schellekens et al., 2016). Six of the models are global hydrological models (GHMs) while four of the models are land surface models (LSMs). GHMs have traditionally been designed to
35 simulate (sub-)surface water fluxes and storages, while LSMs have traditionally been designed to simulate the soil-vegetation-

Table 1. Overview of, to the best of our knowledge, all macro-scale (continental to global) studies evaluating the runoff estimates of multiple models, sorted by region and then publication date. The present study has been added for the sake of completeness.

Study	Region	Number of models	Number of catchments (size range)	Evaluation time scale(s)
Lohmann et al. (2004)	Cont. USA	4	1145 (23 to 10 000 km^2)	Daily, monthly, annual, long term
Xia et al. (2012)	Cont. USA	4	969 (23 to 1 353 280 km^2)	Daily, weekly, monthly, annual, long term
Prudhomme et al. (2011)	Europe	3	579 ($<$ 1000 km^2)	Daily
Gudmundsson et al. (2012a)	Europe	9	426 ($<$ 4000 km^2)	Daily, annual, long term
Gudmundsson et al. (2012b)	Europe	9	426 ($<$ 4000 km^2)	Annual, long term
Greuell et al. (2015)	Europe	5	46 (9948 to 658 340 km^2)	Daily, monthly, annual, long term
Gudmundsson and Seneviratne (2015)	Europe	10	426 ($<$ 4000 km^2)	Monthly, annual, long term
Milly et al. (2005)	Global	12	165 ($>$ 50 000 km^2)	Long term
Decharme and Douville (2006)	Global	6	80 (100 000 to 4 758 000 km^2)	Daily, monthly
Decharme and Douville (2007)	Global	6	80 (100 000 to 4 758 000 km^2)	Monthly
Decharme (2007)	Global	2	80 (100 000 to 4 758 000 km^2)	Monthly
Materia et al. (2010)	Global	13	30 (82 000 to 4 677 000 km^2)	Monthly
Zaitchik et al. (2010)	Global	4	66 (19 000 to 4 600 000 km^2)	Daily, annual
Haddeland et al. (2011)	Global	11	8 (650 000 to 4 600 000 km^2)	Monthly
Zhou et al. (2012)	Global	14	150 (not specified; $\gg$ 10 000 km^2)	Annual
Van Dijk et al. (2013b)	Global	5	6192 (10 to 10 000 km^2)	Monthly
Beck et al. (2015)	Global	4	4079 (10 to 10 000 km^2)	Daily, long term
Yang et al. (2015)	Global	7	16 (135 757 to 3 475 000 km^2)	Monthly, annual
Zhang et al. (2016)	Global	4	644 ($\gg$ 2000 km^2)	Monthly, annual
Beck et al. (2016)	Global	10	1113 (10 to 10 000 km^2)	Daily, 5-day, monthly, long term
This study	Global	10	966 (1000 to 5000 km^2)	Daily, 5-day, monthly, annual, long term

atmosphere interactions within climate models (Haddeland et al., 2011; Bierkens, 2015). GHMs generally represent hydrological processes in a more conceptual way, solve only the water balance, commonly operate at daily time steps, and typically have a small number of soil layers (≤ 3 in the current study) and a single snow layer. Conversely, LSMs generally represent hydrological processes in a more physically-based way, solve both the water and energy balances, typically operate at (sub-)hourly time steps, and tend to have many soil and snow layers (4–11 and 1–12, respectively, in the current study; for more details on the models, see Table 1 of Schellekens et al., 2016). The present study aims to comprehensively evaluate the runoff estimates of these ten models across the globe in an effort to answer the following pertinent research questions:

1. How well do the different models simulate runoff?

2. How well do the models perform in terms of long-term runoff trends?

3. How do the results of the GHMs differ, if at all, from those of the LSMs?

4. Are calibration and regionalization important or even essential?

5. What is the impact of the forcing data on the simulated runoff?

6. How valuable are multi-model ensembles for improving runoff estimates?

7. Do all models show the early bias in runoff timing in snow-dominated catchments previously documented (e.g., Zaitchik et al., 2010) and what is the cause?

We use daily streamflow observations during 1979–2012 from a large, highly diverse, quality-controlled set of medium sized catchments, which allows us to compare the performance among different regions and climate types (Andréassian et al., 2007; Stahl et al., 2011; Gupta et al., 2014). Moreover, we use a broad range of performance metrics, including runoff signatures (measures that quantify the hydrograph shape such as runoff coefficient and baseflow index; Olden and Poff, 2003; Monk et al., 2007) that can be related to specific hydrological processes (Yilmaz et al., 2008).

2   Data

2.1   Forcing

The models were all driven by the daily $0.5°$ WATCH Forcing Data ERA-Interim (WFDEI) meteorological dataset (1979–2012; Weedon et al., 2014) with the precipitation (P) data adjusted using the monthly $0.5°$ gauge-based Climate Research Unit (CRU) TS3.1 dataset (Harris et al., 2013). Although the models all used the same P data, they used potential evaporation (PET) derived using diverse formulations, ranging from the temperature-based Hamon equation (PCR-GLOBWB) to various radiation-based approaches (WaterGAP3, SWBM, and HBV-SIMREG), the Penman-Monteith combination equation (HTES-SEL, JULES, LISFLOOD, SURFEX, and W3RA), and a surface-energy balance approach (ORCHIDEE). The models also used different datasets for non-meteorologic inputs. For more details, see Schellekens et al. (2016).

2.2   Simulated runoff

Table 2 lists the ten state-of-the-art macro-scale hydrological models of which we evaluated the simulated daily unrouted runoff depths (mm d^{-1}). The data used in this study have been named Tier-1 and represent an initial run by all participating modeling groups (Schellekens et al., 2016). All data were acquired through the eartH2Observe Water Cycle Integrator (WCI; http://wci.earth2observe.eu), and for each model the sum of the subsurface and surface runoff components was calculated. Six of the models are GHMs (LISFLOOD, PCR-GLOBWB, SWBM, W3RA, WaterGAP3, and HBV-SIMREG) and four are LSMs (HTESSEL, JULES, ORCHIDEE, and SURFEX). The GHMs were all run at daily time steps and the LSMs at hourly and 15-minute time steps. The models were run at a $0.5°$ spatial resolution, with the exception of LISFLOOD and WaterGAP3, which were run at $0.1°$ and $0.08°$, respectively. For the analysis, however, all model output was resampled to a common $0.5°$ spatial and daily temporal resolution. Four of the models were subjected to varying degrees of calibration to improve their parameters (LISFLOOD, SWBM, WaterGAP3, and HBV-SIMREG; see Section 4.4 for specifics). More details concerning the models can be found in Table 1 of Schellekens et al. (2016).

Table 2. Overview of the hydrological models considered in this study. For definitions of the model name acronyms, see Schellekens et al. (2016). Definitions of model-class acronyms: GHM, global hydrological model; and LSM, land surface model.

Model name	Data provider(s)	Reference(s)	Model class
HTESSEL	European Centre for Medium-Range Weather Forecasts (ECMWF)	Balsamo et al. (2009, 2011)	LSM
JULES	Natural Environment Research Council (NERC)	Best et al. (2011)	LSM
LISFLOOD	Joint Research Centre (JRC)	Burek et al. (2013)	GHM
ORCHIDEE	Centre National de la Recherche Scientifique (CNRS)	Krinner et al. (2005)	LSM
PCR-GLOBWB	University of Utrecht	Van Beek and Bierkens (2009)	GHM
SURFEX	Météo France	Decharme et al. (2011, 2013)	LSM
SWBM	Eidgenössische Technische Hochschule (ETH) Zürich	Orth and Seneviratne (2015)	GHM
W3RA	Australian National University (ANU) and Commonwealth Scientific and Industrial Research Organisation (CSIRO)	Van Dijk (2010)	GHM
WaterGAP3	University of Kassel	Verzano (2009)	GHM
HBV-SIMREG	JRC	Beck et al. (2016)	GHM

2.3 Observed streamflow

Daily and monthly observed streamflow data were used in this study to evaluate the runoff estimates of the models. The observed streamflow and catchment boundary data used in this study originate from the same three sources as Beck et al. (2013, 2015, 2016), namely (i) the Global Runoff Data Centre (GRDC; http://www.bafg.de/GRDC/), (ii) the Geospatial Attributes of Gages for Evaluating Streamflow (GAGES)-II database (Falcone et al., 2010), and (iii) an Australian streamflow data compilation by Peel et al. (2000). The following seven criteria were used to select suitable catchments for our analysis:

1. The streamflow record length was required to be ≥ 5 years (not necessarily consecutive) during 1979–2012 (the temporal span of the simulated runoff data).

2. The catchment area had to be $< 5000\,\mathrm{km}^2$, to minimize the effects of channel routing delays and to reduce the likelihood of significant anthropogenic water use. We could not use larger catchments and evaluate routed streamflow estimates since three of the models did not not simulate river routing (JULES, SWBM, and HBV-SIMREG).

3. The catchment area had to be $> 1000\,\mathrm{km}^2$, to prevent catchments unrepresentative of the 0.5° grid cells (2182 km² at 45°N/S) from confounding the results.

4. To reduce human influences, catchments were required to have $< 2\,\%$ classified as urban (using the "artificial areas" class of the GlobCover version 2.3 map; 300-m resolution; Bontemps et al., 2011) and subject to irrigation (using version 5 of the Global Map of Irrigation Areas—GMIA; 5-min resolution; Siebert et al., 2005).

5. We used the Global Reservoir and Dam (GRanD) database (v1.1; Lehner et al., 2011) to exclude catchments influenced by major reservoirs (defined by total reservoir capacity $> 10\,\%$ of the observed mean annual streamflow).

6. Catchments with forest gain or loss > 20 % of the catchment area (the threshold at which changes in runoff can generally be detected; Bosch and Hewlett, 1982) were excluded using version 1.1 of the Landsat-based forest change dataset (30-m resolution; Hansen et al., 2013).

7. To further reduce the number of disinformative catchments, all streamflow records were visually screened for artifacts and anthropogenic influences (caused by, for example, diversions and impoundments). Furthermore, USA catchments flagged as "non-reference" in the GAGES-II database were discarded, and GRDC catchments for which the catchment boundaries could not be reliably determined were discarded (Lehner, 2012).

In total 966 catchments (median size 1970 km^2; median record length 19 y during 1979–2012) were found to be suitable for the analysis, of which 641 catchments have daily streamflow data and 325 catchments (mainly located in Russia) have only monthly streamflow data. The locations of the selected catchments will be shown in the Results section. All observed streamflow data were converted to runoff in mm d^{-1} using the provided catchment areas.

3 Methodology

3.1 Model evaluation

The simulated runoff of the models were evaluated in five ways. First, for each catchment, we calculated the differences D $(-)$ between simulated and observed values of several runoff signatures. Table 3 lists the six runoff signatures selected including their computation from the period with simultaneous simulated and observed runoff. The baseflow index (BFI), square-root transformed 1st percentile exceedance flow (Q1), and square-root transformed 99th percentile exceedance flow (Q99) require daily (rather than monthly) flow data. To compute the flow timing (T50) from monthly data, we first computed daily time series from monthly time series using linear interpolation. The square-root transformed runoff coefficient (RC), square-root transformed mean annual flow (MAR), Q1, and Q99 values were square-root transformed to give more weight to small values. D was computed according to:

$$D_q = \frac{Y_{q\,\mathrm{sim}} - Y_{q\,\mathrm{obs}}}{\sigma_q}, \tag{1}$$

where Y represent the values of the runoff signatures $(-)$, the q subscript denotes the runoff signature, and the 'sim' and 'obs' subscripts refer to simulated and observed, respectively. The σ values $(-)$ are constants that represent the spatial variability in the runoff signatures across the landscape and are used to normalize the D values (i.e., to make the D values of the different signatures intercomparable; see Table 3). The σ values were computed by taking the standard deviation of the observed values. Next, the mean D value over all catchments was computed (expressed by $\overline{D}$). D and $\overline{D}$ values closer to zero correspond to better model performance (see Table 4). It should be noted that, although $\overline{D}$ provides a valuable estimate of the overall performance, a good $\overline{D}$ value may reflect an overestimation in one region that is compensated by an underestimation in another region.

Table 3. The long-term runoff behavioral signatures considered for evaluating the model performance. The signatures were computed, for each catchment, from the entire record of simultaneous observed and simulated runoff. The σ values represent the spatial variability in the runoff signatures across the landscape.

Runoff signature	Units	Description	Evaluated flow aspect	Standard deviation (σ)
RC	–	Square-root transformed runoff coefficient, ratio of long-term runoff to P	Water balance	0.33
MAR	$\sqrt{mm\ yr^{-1}}$	Square-root transformed long-term mean annual runoff	Water balance	11.21
T50	d	The day of the water year marking the timing of the center of mass of flow (Stewart et al., 2005). A water year is defined as the 12-month period from October to September in the Northern Hemisphere and April to March in the Southern Hemisphere	Seasonal flow distribution	34.36
BFI	–	Base flow index, the ratio of long-term baseflow to total runoff; the baseflow portion of the total runoff was computed following the procedure of Gustard et al. (1992), which takes the minima at five-day non-overlapping intervals and subsequently connects the valleys in this series of minima to generate baseflow	Partitioning between quickflow and baseflow, flow peakiness	0.18
Q1	$\sqrt{mm\ d^{-1}}$	Square-root transformed 1st percentile exceedance flow	Peak-flow magnitude	1.27
Q99	$\sqrt{mm\ d^{-1}}$	Square-root transformed 99th percentile exceedance flow	Low-flow magnitude	0.21

Second, to evaluate the temporal variability of the simulated runoff time series, we computed Pearson linear correlation coefficients (r) between daily, log-transformed daily, 5-day, monthly, monthly climatic, and annual time series of simulated and observed runoff (termed r_{dly}, $r_{\mathrm{dly\,log}}$, $r_{\mathrm{5\,day}}$, r_{mon}, $r_{\mathrm{mon\,clim}}$, and r_{yr}, respectively). The r_{dly}, $r_{\mathrm{dly\,log}}$, and $r_{\mathrm{5\,day}}$ values were only computed for catchments with daily observations. If monthly data were not supplied by the data providers, monthly values were computed by simple averaging of the daily data only if > 25 non-missing values were available. Annual values were computed by simple averaging of the monthly data (either supplied or computed) only if > 10 non-missing values were available. We subsequently computed for each model and metric the mean r value over all catchments, expressed by $\overline{r}$. The r and $\overline{r}$ values range from -1 to 1, with higher values corresponding to better model performance (see Table 4).

Third, to summarize the overall performance of each model, we computed for each catchment a summary performance statistic (termed OS) incorporating the previously mentioned metrics, and computed the mean value over all catchments ($\overline{\mathrm{OS}}$). The OS consists of two parts, of which the first ($\mathrm{OS}_{\mathrm{sig}}$) considers the performance in terms of runoff signatures and is defined as:

$$\mathrm{OS}_{\mathrm{sig}} = 1 - \mathrm{mean}\left[|D_{\mathrm{RC}}|, |D_{\mathrm{MAR}}|, |D_{\mathrm{T50}}|, |D_{\mathrm{BFI}}|, |D_{\mathrm{Q1}}|, |D_{\mathrm{Q99}}|\right]. \tag{2}$$

The second part ($\mathrm{OS}_{\mathrm{var}}$) evaluates the performance in terms of temporal variability, and is defined as:

$$\mathrm{OS}_{\mathrm{var}} = \mathrm{mean}\left[r_{\mathrm{dly}}, r_{\mathrm{dly\,log}}, r_{\mathrm{5\,day}}, r_{\mathrm{mon}}, r_{\mathrm{mon\,clim}}, r_{\mathrm{yr}}\right]. \tag{3}$$

The summary score is subsequently computed following:

$$\mathrm{OS} = \frac{\mathrm{OS}_{\mathrm{sig}} + \mathrm{OS}_{\mathrm{var}}}{2}. \tag{4}$$

Table 4. Qualitative descriptions of intervals of the performance metrics to aid in interpreting the results.

	$|D|$	r, ρ	OS
Excellent	$[0, 0.2)$	$[0.8, 1]$	$[0.8, 1]$
Good	$[0.2, 0.4)$	$[0.6, 0.8)$	$[0.6, 0.8)$
Moderate	$[0.4, 0.6)$	$[0.4, 0.6)$	$[0.4, 0.6)$
Fair	$[0.6, 0.8)$	$[0.2, 0.4)$	$[0.2, 0.4)$
Poor	$[0.8, +\infty]$	$[-1, 0.2)$	$[-\infty, 0.2)$

The BFI, Q1, and Q99 components of Equation 2 and the r_{dly} and $r_{\mathrm{dly\,log}}$ components of Equation 3 were omitted if daily observations were unavailable for a particular catchment. Higher OS values correspond to better model performance; the maximum attainable value is 1 (see Table 4).

Fourth, to evaluate the ability of each model to simulate the variability among the catchments in the six previously men-
5 tioned runoff signatures, Spearman rank correlation coefficients (ρ) were computed between simulated and observed values of the runoff signatures. Spearman rank correlation coefficients rather than Pearson linear correlation coefficients were used to minimize the influence of outliers. The ρ values range from -1 to 1, with higher values corresponding to better model performance (see Table 4).

Fifth, we computed trends in simulated and observed mean annual runoff time series (termed MAR trend) using the sim-
10 ple non-parametric approach of Sen (1968). We subsequently calculated the ρ between simulated and observed MAR trends ($\rho_{\mathrm{MAR\,trend}}$), reflecting the agreement in spatial trend patterns.

Sixth and last, we produced density plots of grid cell values of aridity index (AI; ratio of long-term available energy to P) versus runoff coefficient (RC; ratio of long-term simulated runoff to P), revealing how the models behave in terms of RC under different climatic conditions. To estimate the available energy we used PET for four models (ORCHIDEE, PCR-GLOBWB,
15 W3RA, and WaterGAP3) and net radiation for three models (HTESSEL, JULES, and SURFEX). For the remaining models estimates of the available energy were not available.

For the evaluation, we used for each catchment the simulated runoff time series of the $0.5°$ grid cell with its center located within the catchment. However, if multiple grid cell centers were located within the catchment, we calculated the mean simulated runoff time series, and if no grid cell center was located within the catchment, we used the simulated runoff time series
20 of the grid cell with its center located closest to the catchment centroid.

3.2 Multi-model ensembles

Ensemble modeling—using the outputs from multiple models or from different realizations of the same model—typically improves predictive accuracy and is widely used in atmospheric, climate, and hydrological sciences (Wandishin et al., 2001; Tebaldi and Knutti, 2007; Breuer et al., 2009; Viney et al., 2009). We tested two ways of combining the runoff estimates
25 of the individual models into ensembles. First, for each $0.5°$ grid cell and day with non-missing values for all models, the mean simulated runoff of the ten models was calculated (i.e., equal weights were assigned to the models). The resulting runoff

estimates will be referred to hereafter as "MEAN-All". Second, we computed the mean based on only the four models that performed best in terms of $\overline{\text{OS}}$, to examine the effect of excluding less accurate models. These runoff estimates will be referred to hereafter as "MEAN-Best4".

3.3 Caveats

5 There are a number of caveats that should be kept in mind when interpreting the results. First, some of the models (notably the LSMs) were not traditionally developed to estimate daily runoff for such small catchments. Some of the GHMs, on the other hand, have runoff estimation in small catchments among their primary aims (e.g., LISFLOOD, WaterGAP3, W3RA, and HBV-SIMREG), and four GHMs were even explicitly calibrated against observations (LISFLOOD, SWBM, WaterGAP3, and HBV-SIMREG; see Section 4.4 for specifics). Second, a model performing poorly in one respect may well perform better for

10 other hydrological variables, climates, catchments, or performance metrics. Third, a poor model performance could simply be the result of suboptimal parameter values. Fourth, some studies have found that less accurate models may still lead to a better ensemble mean (Ajami et al., 2006; Viney et al., 2009), although this did not appear to be the case here (see Section 4.6). Fifth, we stress that while some models may perform well, they are inherently unsuitable for specific types of impact assessments. For example, SWBM and HBV-SIMREG do not account for physical differences among land-cover types and hence cannot be

15 used for studies assessing the hydrological impacts of changes in land cover. Sixth and finally, the forcing data quality has an important influence on the evaluation results that should not be overlooked.

4   Results and discussion

In this section we will answer the questions posed in the introduction.

4.1   How well do the different models simulate runoff?

20 Table 5 show, for the uncalibrated models, the calibrated models, and the ensembles, (i) the mean difference between simulated and observed values of the (normalized) runoff signatures ($\overline{D}$), (ii) the mean temporal correlation between simulated and observed runoff time series ($\overline{r}$), and (iii) the mean overall performance in terms of runoff signatures and temporal correlation coefficients ($\overline{\text{OS}}$). HTESSEL obtained negative D values for the square-root transformed runoff coefficient (RC) and the square-root transformed mean annual runoff (MAR), indicating it underestimates runoff. JULES performed moderately in

25 terms of temporal correlation, as indicated by the low r values. Conversely, LISFLOOD performed good overall, particularly in terms of temporal correlation, although it tends to overestimate RC and MAR. ORCHIDEE appears to strongly underestimate runoff and performed fairly in terms of temporal correlation, whereas PCR-GLOBWB shows moderate to good scores for most metrics. Apart from a much too early bias in the flow timing (T50), SURFEX demonstrated moderate to good performance overall. Similar to SURFEX, W3RA exhibits a very early bias in T50, but generally obtained moderate to good scores.

30 WaterGAP3 and particularly HBV-SIMREG outperformed the other models in most cases. JULES, ORCHIDEE, SURFEX, WaterGAP3, and especially SWBM displayed negative D values for the baseflow index (BFI) and the square-root transformed

99th flow percentile (Q99), and a positive D value for the square-root transformed 1st flow percentile (Q1; Table 5), suggesting they consistently overestimate quickflow. Conversely, LISFLOOD and particularly PCR-GLOBWB exhibited positive D values for BFI and Q99, and a negative D value for Q1, indicating they tend to underestimate quickflow.

Table 5 also presents, for the ten models and the ensembles, the spatial correlation between simulated and observed values of the runoff signatures (ρ). HTESSEL, JULES, W3RA, WaterGAP3, and HBV-SIMREG performed good overall, while the remaining models performed moderately overall. PCR-GLOBWB, SURFEX, and WaterGAP3 performed poorly in terms of BFI, while SWBM obtained a poor score for Q99. WaterGAP3 performed good to excellent for all signatures except BFI, likely due to the empirical estimation of groundwater recharge and thus baseflow as a function of landscape characteristics (Döll and Fiedler, 2008). HBV-SIMREG attained good to excellent ρ values for all signatures. The models generally performed best for T50 and worst for BFI among the signatures.

Table 5 also shows, for the ten models and the ensembles, $\overline{\text{OS}}$ scores for the major Köppen-Geiger climate types. We used the newly produced Köppen-Geiger climate map from Beck et al. (2016) which is based on the high-quality WorldClim climatic dataset (Hijmans et al., 2005) supplemented with regional climatic datasets for the USA (Daly et al., 1994) and New Zealand (Tait et al., 2006). All four LSMs (HTESSEL, JULES, ORCHIDEE, and SURFEX) generally demonstrated fair performance in cold and polar climates. Conversely, PCR-GLOBWB demonstrated poor performance in tropical and arid climates, likely due to the overestimation of baseflow. SWBM performed moderately only in arid catchments, probably at least partly due to the lack of baseflow under these conditions (Pilgrim et al., 1988; Beck et al., 2013). Similarly, Orth et al. (2015) found that SWBM performs well during dry periods for eight small Swiss catchments (60–392 km^2). Only LISFLOOD WaterGAP3 and HBV-SIMREG exhibited at least moderate performance for all climates.

Figure 1 presents, for the ten models and the ensembles, maps of simulated minus observed MAR for the catchments, revealing the data underlying the MAR $\overline{D}$ and ρ values listed in Table 5. Maps of all other runoff signatures are presented in Supplementary material Figures S1.2–8. HTESSEL and ORCHIDEE strongly underestimate runoff for most of the catchments, while LISFLOOD appears to strongly overestimate runoff for most of the globe with the exception of snow-dominated regions. All models showed negative MAR biases in snow-dominated regions such as Alaska, the Rocky Mountains, and southern Russia, while they consistently showed positive MAR biases for the Great Plains (USA) and southern Australia. Figure 2 shows, for the ten models and the ensembles, maps of the correlation between simulated and observed monthly flows (r_{mon}) for the catchments, showing the data underlying the $\overline{r_{\text{mon}}}$ values presented in Table 5. Maps of all other temporal variability metrics are presented in Supplementary material Figures S1.9–14. In general, the GHMs obtained good r_{mon} values for most catchments, while the LSMs obtained moderate r_{mon} values for most catchments. All LSMs showed poor to fair r_{mon} values for snow-dominated catchments.

Although the NSE has been widely criticized for being overly sensitive to the magnitude and timing of peak flows (e.g., Schaefli and Gupta, 2007; Jain and Sudheer, 2008; Criss and Winston, 2008; Gupta et al., 2009), we did calculate NSE scores to allow the present results to be put in the context of previous macro-scale studies (see Supplementary material Table S1). For most models negative median NSE scores were obtained, similar to Zhang et al. (2016), who evaluated the monthly and annual runoff estimates from 14 (uncalibrated) macro-scale models in 644 large Australian catchments (> 2000 km^2). Our

Table 5. For the individual models and the ensembles, (i) the mean difference between simulated and observed values of the (normalized) runoff signatures ($\overline{D}$), (ii) the mean temporal correlation between simulated and observed runoff time series ($\overline{\tau}$), (iii) the mean overall performance in terms of runoff signatures and temporal correlation ($\overline{OS}$), and (iv) the spatial correlation between simulated and observed values of the runoff signatures (ρ). Color legend: poor, fair, moderate, good, excellent (see Table 4 for the corresponding metric intervals). See Figure 1 for the locations of the catchments.

| Metric | Uncalibrated models | | | | | | Calibrated models | | | | Ensembles | |
	HTESSEL	JULES	ORCHIDEE	PCR-GLOBWB	SURFEX	W3RA	LISFLOOD	SWBM	WaterGAP3	HBV-SIMREG	MEAN-All	MEAN-Best4
(i) Mean difference between simulated and observed values of the (normalized) runoff signatures												
$\overline{D}_{RC}$ (n = 966)	−0.47	−0.18	−0.60	−0.02	−0.30	−0.24	0.09	−0.16	−0.10	−0.07	−0.16	−0.05
$\overline{D}_{MAR}$ (n = 966)	−0.30	−0.10	−0.39	0.00	−0.19	−0.13	0.08	−0.07	−0.08	−0.03	−0.09	−0.02
$\overline{D}_{T50}$ (n = 966)	−0.24	−0.35	0.05	−0.18	−0.57	−0.45	−0.08	−0.21	−0.23	−0.02	−0.21	−0.16
$\overline{D}_{BFI}$ (n = 619)	−0.04	−0.84	−0.92	1.02	−1.03	0.23	0.42	−2.12	−0.69	−0.08	−0.12	0.15
$\overline{D}_{Q1}$ (n = 641)	−0.06	0.22	0.17	−0.24	0.31	−0.07	−0.02	0.63	0.31	0.10	0.01	0.00
$\overline{D}_{Q99}$ (n = 641)	−0.17	−0.55	−0.70	0.21	−0.67	0.06	0.27	−1.06	−0.13	−0.02	0.11	0.25
(ii) Mean temporal correlation between simulated and observed runoff time series												
$\overline{r}_{dly}$ (n = 641)	0.33	0.23	0.21	0.34	0.31	0.44	0.59	0.32	0.33	0.56	0.44	0.54
$\overline{r}_{dly\,log}$ (n = 641)	0.50	0.41	0.33	0.50	0.51	0.56	0.70	0.34	0.56	0.71	0.64	0.71
$\overline{r}_{5\,day}$ (n = 641)	0.45	0.36	0.33	0.44	0.41	0.52	0.64	0.48	0.52	0.65	0.59	0.65
$\overline{r}_{mon}$ (n = 966)	0.53	0.44	0.40	0.58	0.43	0.57	0.71	0.63	0.65	0.74	0.69	0.72
$\overline{r}_{mon\,clim}$ (n = 966)	0.66	0.50	0.49	0.73	0.47	0.64	0.84	0.75	0.76	0.86	0.80	0.84
$\overline{r}_{yr}$ (n = 966)	0.58	0.61	0.51	0.58	0.57	0.63	0.62	0.60	0.59	0.62	0.64	0.63
(iii) Mean overall performance in terms of runoff signatures and temporal correlation												
All (n = 966)	0.43	0.39	0.26	0.41	0.32	0.46	0.55	0.34	0.52	0.62	0.57	0.60
A: tropical (n = 57)	0.41	0.46	0.28	0.03	0.41	0.39	0.43	0.29	0.40	0.47	0.48	0.47
B: arid (n = 38)	0.52	0.50	0.38	0.07	0.46	0.50	0.32	0.44	0.42	0.55	0.50	0.44
C: temperate (n = 203)	0.46	0.54	0.35	0.37	0.51	0.51	0.52	0.31	0.48	0.61	0.59	0.58
D: cold (n = 633)	0.43	0.34	0.23	0.47	0.25	0.45	0.58	0.35	0.55	0.65	0.59	0.63
E: polar (n = 35)	0.32	0.25	0.20	0.53	0.23	0.33	0.60	0.25	0.44	0.60	0.51	0.57
(iv) Spatial correlation between simulated and observed values of the runoff signatures												
ρ_{RC} (n = 966)	0.67	0.64	0.30	0.56	0.65	0.60	0.57	0.54	0.82	0.70	0.72	0.79
ρ_{MAR} (n = 966)	0.80	0.78	0.61	0.73	0.79	0.77	0.71	0.74	0.87	0.81	0.81	0.83
ρ_{T50} (n = 966)	0.76	0.82	0.66	0.63	0.78	0.85	0.87	0.88	0.88	0.91	0.91	0.90
ρ_{BFI} (n = 619)	0.38	0.28	0.46	0.10	0.01	0.35	0.28	0.37	−0.03	0.71	0.55	0.54
ρ_{Q1} (n = 641)	0.77	0.74	0.54	0.53	0.64	0.67	0.65	0.73	0.80	0.76	0.76	0.78
ρ_{Q99} (n = 641)	0.70	0.69	0.51	0.43	0.59	0.68	0.58	0.09	0.71	0.76	0.75	0.74
$\rho_{MAR\,trend}$ (n = 966)	0.37	0.38	0.37	0.36	0.32	0.38	0.42	0.39	0.35	0.37	0.40	0.39
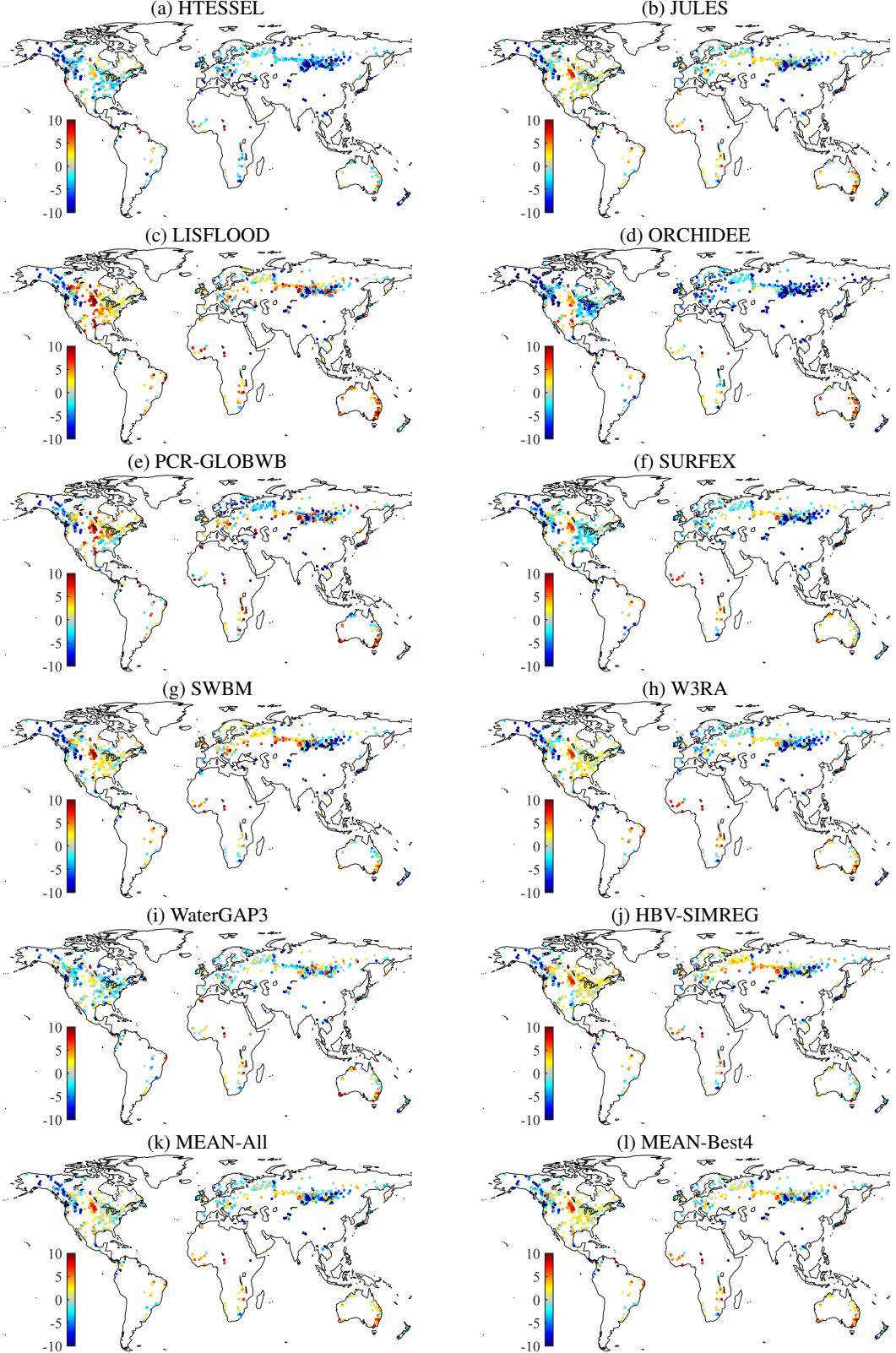
Figure 1. Simulated minus observed square-root transformed mean annual runoff (MAR; units $\sqrt{\text{mm yr}^{-1}}$) for the catchments. Each data point represents a catchment centroid ($n = 966$). Red (blue) indicates an overestimated (underestimated) MAR relative to the observations.
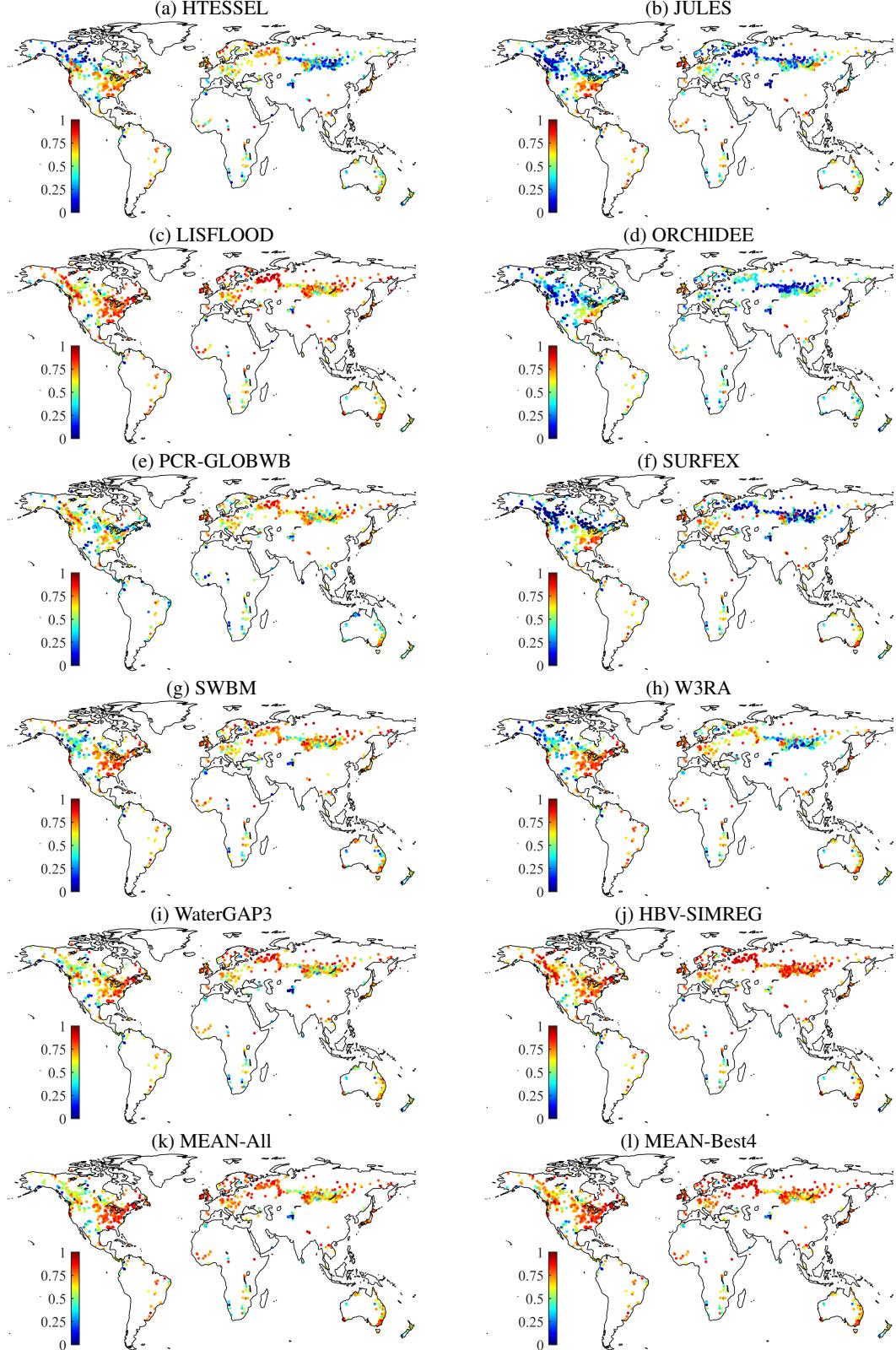
Figure 2. Correlation coefficients calculated between simulated and observed monthly runoff (r_{mon}; unitless) for the catchments. Each data point represents a catchment centroid ($n = 966$).

scores are, however, slightly lower than those obtained by Lohmann et al. (2004) and Xia et al. (2012), who evaluated the daily runoff estimates from four (uncalibrated) macro-scale models in about a thousand small-to-medium sized USA catchments ($< 10\,000$ km^2), but this is probably attributable to the high quality of the USA forcing data (**?**). They are also somewhat lower than those obtained by Decharme and Douville (2007), who evaluated two (uncalibrated) macro-scale models in 80 large

5   catchments ($> 100\,000$ km^2) around the globe, but this can be explained by their much larger catchment sizes.

Figure 3 shows, for the seven models with data on energy availability, density plots of grid cell values of aridity index (AI; ratio of long-term energy availability to P) versus runoff coefficient (RC; ratio of long-term mean runoff to P), revealing how the models respond in terms of RC to different climatic conditions. Also shown are the energy-limit line for which actual evaporation equals the available energy, the water-limit line for which runoff equals P, and the Budyko (1974) curve, the most

10 well-known among several similar empirical relationships describing the competition between runoff and actual evaporation (Ol'dekop, 1911; Pike, 1964; Zhang et al., 2001; Porporato et al., 2004). Given its empirical nature, the Budyko curve should only be used for visual reference, and not to judge the performance of the different models. Besides the striking differences in behavior among the models, it can be seen that HTESSEL, JULES, PCR-GLOBWB, W3RA, and WaterGAP3 do not adhere to the water and/or energy limits (Figures 3a, 3b, 3d, 3f, and 3g, respectively). For WaterGAP3, this may be due to the use of

15 calibration factors, which have the potential to generate runoff that can go beyond the physical limits in an effort to compensate for errors in the P, PET, or streamflow data. For the other models this could be indicative of issues with the runoff and/or evaporation routines. The larger spread found for the models for which we used net radiation to estimate the available energy (HTESSEL, JULES, and SURFEX; Figures 3a, 3b, and 3e, respectively) is because the majority of the net radiation is converted to sensible heat rather than latent heat in cold climates (Kleidon et al., 2014).

20 It is generally difficult to gain insight into why a particular model performs as it does due to the large number of interacting model components, equations, and parameters. Nevertheless, the underestimation of runoff by HTESSEL probably reflects the excessive evaporation by HTESSEL previously reported by Haddeland et al. (2011). PCR-GLOBWB most likely suffers from suboptimal baseflow-related parameter values, since its structure is similar to that of LISFLOOD which performs markedly better. SWBM clearly suffers from the absence of a baseflow routine outside (semi-)arid regions. Although W3RA

25 and HBV-SIMREG use an identical snow routine, W3RA performs considerably worse in snow-dominated regions, probably because HBV-SIMREG uses a snowfall gauge undercatch correction factor. The unsatisfactory performance demonstrated by the LSMs in snow-dominated regions could be related to deficiencies in the snow routines or the energy balance estimates (see Section 4.3). WaterGAP3 and particularly HBV-SIMREG performed quite well overall, likely because of their comprehensive calibration (see Section 4.4). In any case, the pronounced inter-model performance spread found here suggests that model

30 choice should be regarded as a critical step in any hydrological modeling study. Moreover, it underscores the importance of hydrological model uncertainty in addition to climate input uncertainty, as also emphasized in several other recent macro-scale studies (Haddeland et al., 2011; Schewe et al., 2013; Prudhomme et al., 2014; Mendoza et al., 2015b; Giuntoli et al., 2015a). Currently, the large majority of studies assessing the hydrological impacts of climate change completely neglect hydrological model uncertainty (Teutschbein and Seibert, 2010).
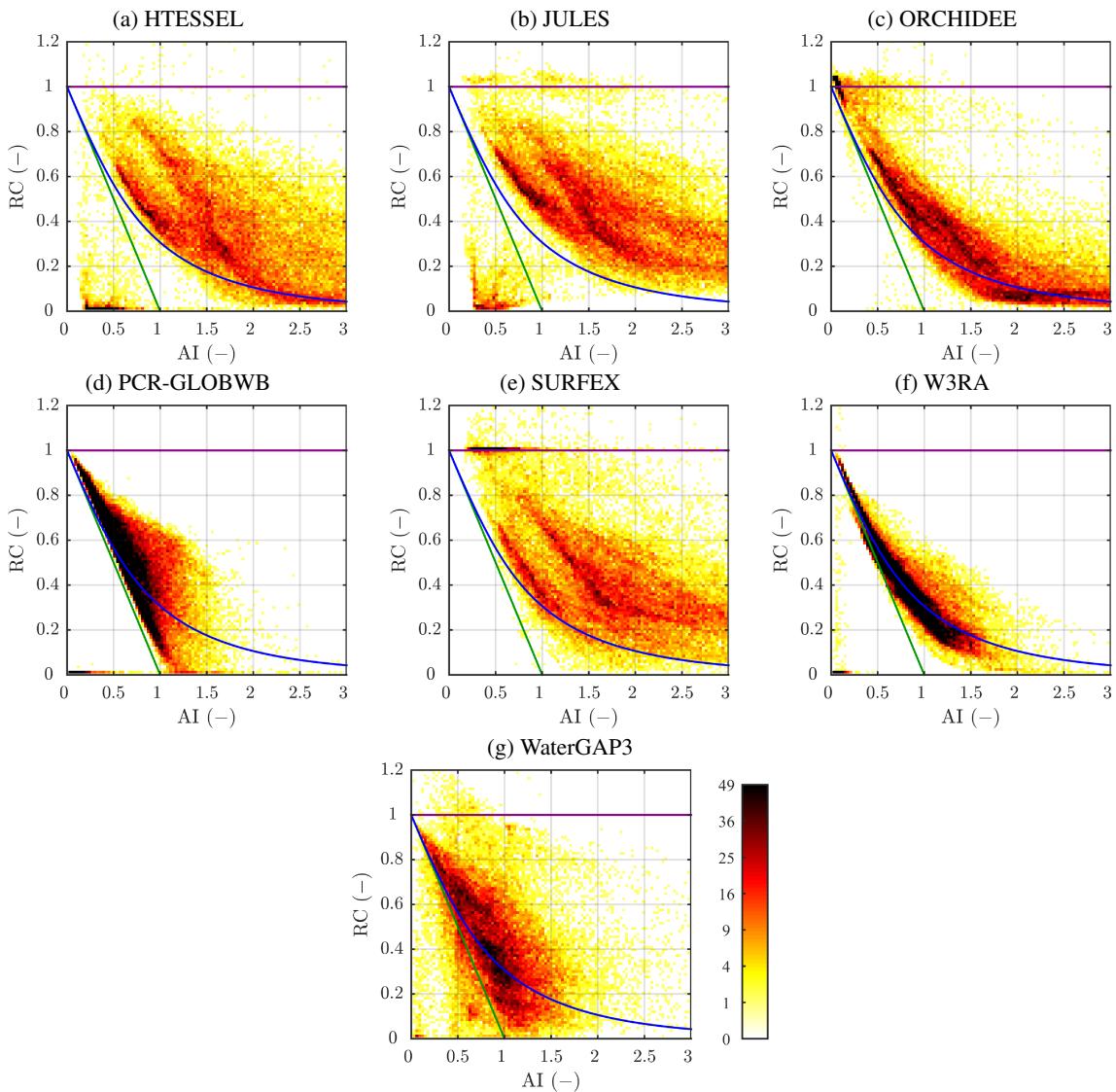
Figure 3. For the seven models with data on the available energy, density plots of grid cell values of aridity index (AI) versus runoff coefficient (RC). Grid cells $> 50°$N/S were excluded from the analysis. The green line represents the energy limit for which actual evaporation equals PET, the purple line represents the water limit for which runoff equals P, whereas the blue line represents the Budyko (1974) curve.

4.2 How well do the models perform in terms of long-term runoff trends?

The models displayed very similar MAR trends (Supplementary material Figure S1.8), meaning they respond similarly to climate variability, given that none of the models account for land-use or land cover changes, urbanization, reservoir construction, or increasing atmospheric CO_2. However, the models obtained rather low spatial (Spearman) correlation coefficients ($\rho_{\mathrm{MAR\,trend}}$) ranging from 0.32 (SURFEX) to 0.42 (LISFLOOD; Table 5), indicating that the simulated MAR trends correspond fairly to moderately well to the observed ones. These values are lower than the (Pearson) correlation coefficients ranging from 0.52 to 0.63 obtained by Stahl et al. (2012), who evaluated MAR trends from seven models using observations from 293 small European catchments (100–1000 km^2), presumably due to the better quality of the European meteorological forcing and observed streamflow data. Milly et al. (2005) evaluated MAR trends from a 12-model ensemble using observations from 165 large catchments ($> 50\,000$ km^2) around the globe, obtaining a (Pearson) correlation coefficient of 0.34 which is similar to ours. These low correlations, which were somewhat unexpected given the relative ease with which MAR can be estimated (e.g., Westerberg and McMillan, 2015; Beck et al., 2015), may be indicative of changes in non-climatic drivers of hydrological change or drift errors in the forcing or observed streamflow data. We expect the inter-model variability in trends to be higher and the agreement with observations to be even lower for seasonal and monthly averages as well as runoff signatures sensitive to the shape of individual flow events (cf. Bastola et al., 2011; Gosling et al., 2011). Overall, these results suggest that studies using global-scale datasets to assess the impacts of past climate change on runoff in small-to-medium sized catchments should be interpreted with considerable caution.

4.3 How do the results of the GHMs differ, if at all, from those of the LSMs?

Similar to Haddeland et al. (2011), the LSMs were found to produce less runoff overall (Table 5 and Figure 1), perhaps due to their use of physically-based Richards-Darcy type equations which neglect preferential flows. We further found that the GHMs perform, on average, worse than the LSMs in rain-dominated regions: the GHMs (excluding the comprehensively calibrated models—WaterGAP3 and HBV-SIMREG; see Section 4.4) obtained mean $\overline{\mathrm{OS}}$ scores of 0.28, 0.33, and 0.43 for tropical, arid, and temperate climates, respectively, while the same values for the LSMs are 0.39, 0.47, and 0.47, respectively (Table 5). However, the lower performance of the GHMs is primarily attributable to PCR-GLOBWB and SWBM. As mentioned before, PCR-GLOBWB probably suffers from a suboptimal baseflow-related parameterization, while SWBM suffers from the absence of a baseflow routine.

The GHMs do appear to perform consistently better than the LSMs in snow-dominated regions: the GHMs (again excluding WaterGAP3 and HBV-SIMREG) obtained mean $\overline{\mathrm{OS}}$ scores of 0.46 and 0.32 for cold and polar climates, respectively, while the same values for the LSMs are 0.31 and 0.25, respectively (Table 5). The performance of the LSMs appears to be mainly due to a very early bias in flow timing, a very low baseflow contribution, and a misrepresentation of the seasonal cycle (Supplementary material Figures S1.4, S1.5, and S1.13, respectively). Our results are in agreement with Giuntoli et al. (2015b), who found five GHMs to outperform, on average, four LSMs using observations from 252 temperate and cold catchments (64 to 1 350 000 km^2) located in the central USA, and with Zhang et al. (2016), who found that two LSMs performed considerably worse than

two GHMs in cold and polar regions using observations from 644 catchments (> 2000 km^2, upper limit not reported) around the globe. The poorer performance obtained by the LSMs is probably indicative of differences between the snow routines used by GHMs and LSMs. The GHMs use relatively simple conceptual temperature-index snow routines driven by air temperature which can be estimated with relative ease, whereas the LSMs use more complex physically-based energy balance snow routines driven by estimates of energy balance components which are subject to considerable uncertainty, particularly in regions with complex topography (Ferguson, 1999). Although several previous studies have found that the two types of snow routines yield comparable performance (e.g., WMO, 1986; Franz et al., 2008; Zeinivand and De Smedt, 2009; Debele et al., 2010), these studies used a very small number of relatively well-instrumented catchments (six, two, one, and three, respectively) which may have led to less-generalizable conclusions. Overall, it appears that the energy balance estimates and snow routines used by the LSMs require re-evaluation (cf. Zhang et al., 2016).

4.4 Are calibration and regionalization important or even essential?

Calibration is important for both conceptual and physically-based hydrological models to provide more accurate runoff estimates, to account for (i) the impossibility of measuring all required model parameters at the model application scale, (ii) lack of process understanding, (iii) possibly overly simplistic process representations, (iv) the spatio-temporal discretization of highly heterogeneous rainfall-runoff processes, and (v) errors in the forcing data (Beven, 1989; Blöschl and Sivapalan, 1995; Duan et al., 2001, 2006; McDonnell et al., 2007; Nasonova et al., 2009; Rosero et al., 2011; Minville et al., 2014). Yet, despite the development of numerous calibration techniques over the last 50 years (Dawdy and O'Donnell, 1965; Duan et al., 2004) and the current widespread availability of streamflow observations (Hannah et al., 2011), macro-scale models generally tend to be uncalibrated (Sooda and Smakhtin, 2015; Bierkens, 2015; Kauffeldt et al., 2016). This is perhaps mainly due to (i) the substantial amount of work involved with calibration (e.g., Bock et al., 2015), (ii) the risk of obtaining unrealistic parameters due to equifinality and data issues (Andréassian et al., 2012), and (iii) the lack of a commonly accepted regionalization technique (Beck et al., 2016). In addition, the modeler may feel that since their model is physically based, it does not require calibration (Beven, 1989). LSMs in particular are rarely calibrated against runoff, likely because: (i) runoff estimation is generally not among the primary aims of LSMs; (ii) for water transport in the soil, LSMs typically use Richards-Darcy type equations which are computationally expensive and require a fine vertical and temporal soil discretization; and (iii) LSMs often do not account for river routing, confounding the calibration of large catchments. Instead, the parameters in macro-scale models are usually based on "expert opinion" and thus founded on the bold assumption that the modeler sufficiently understands the hydrological processes, feedbacks, and parameter interactions taking place within the model for any location on Earth.

Nevertheless, out of the ten models considered in this study, four use parameters derived by calibration (LISFLOOD, SWBM, WaterGAP3, and HBV-SIMREG—all GHMs). LISFLOOD was calibrated against observed streamflow for 24 large catchments ($84\,230$ to $4\,680\,000$ km^2) across the globe using the WFDEI forcing and an aggregate objective function incorporating bias, NSE, and log-transformed NSE computed from daily streamflow data. The calibration might have influenced the present evaluation; although we used much smaller catchments (1000 to 5000 km^2), 47 % of our catchments are located within the calibration catchments. SWBM uses a spatially-uniform parameter set based on calibration using the E-OBS forcing (Hay-

lock et al., 2008) against European data on such key hydrologic variables as soil moisture, total water storage, evaporation, and runoff (Orth and Seneviratne, 2015). For the calibration against runoff, they used observations from 436 small European catchments (mostly < 1000 km^2), and considered daily and monthly correlations as well as bias. The calibrated parameter set was subsequently applied globally. Besides the addition of a baseflow routine, SWBM would probably benefit from region-

5 alized parameters that vary according to landscape characteristics. WaterGAP3 has been calibrated using the WFDEI forcing in terms of bias for the interstation regions (the catchment of a station excluding the catchments of nested upstream stations) of 2071 stations (catchment size ranging from 2830 to 966 321 km^2) around the globe, some of which have also been used in the current evaluation. The calibrated parameters were subsequently regionalized to ungauged regions using multiple linear regression based on six predictors (Döll et al., 2003). The model does indeed perform very well for MAR and thus RC, but this

10 did not necessarily translate into good performance for BFI (Table 5, and Figures 1 and 2). HBV-SIMREG also uses region-alized parameter fields, produced by transferring calibrated parameters from 674 small-to-medium sized "donor" catchments (10 to 10 000 km^2) across the globe to "receptor" grid cells with similar climatic and physiographic characteristics (Beck et al., 2016). In their study, Beck et al. (2016) show that HBV using spatially-uniform parameters performs within the range of the other models, confirming that the relatively good performance of HBV-SIMREG stems from the regionalization exercise. In

15 addition, although Beck et al. (2016) did not use the WFDEI forcing for the calibration, they calibrated against several of the performance metrics also used here and used 179 of our catchments as parameter donors, further explaining the relatively good performance obtained by HBV-SIMREG (Table 5, and Figures 1 and 2).

Overall, it appears that the calibration exercises for WaterGAP3, HBV-SIMREG, and possibly LISFLOOD have resulted in markedly improved performance. However, WaterGAP3 performed poorly in terms of ρ_{BFI} (Table 5), meaning the calibration

20 of MAR did not translate into better BFI performance. These results underscore the benefits of calibrated parameters over *a priori* parameters (cf. Duan et al., 2006; Hunger and Döll, 2008; Nasonova et al., 2009; Rosero et al., 2011; Greuell et al., 2015; Zhang et al., 2016) and highlight the importance of using an objective function for the calibration that incorporates a broad range of metrics related to various important aspects of the hydrograph (cf. Gupta et al., 2008; Vis et al., 2015; Shafii and Tolson, 2015). These results also emphasize the usefulness of regionalization techniques (Parajka et al., 2013), which typically

25 enhance performance over the entire model domain and are thus of particular value for macro-scale modeling, given that the majority of the land surface is ungauged or poorly gauged (Sivapalan, 2003; Hannah et al., 2011). However, although there are numerous studies performing regionalization at a regional scale (see reviews by He et al., 2011; Hrachowitz et al., 2013; Razavi and Coulibaly, 2013; Parajka et al., 2013), only few studies have attempted regionalization at a macro scale (see review by Beck et al., 2016). We argue that more effort should be devoted to regionalizing the parameters of macro-scale models (cf.

30 Bierkens, 2015; Döll et al., 2015).

It should be noted, however, that the potential performance improvement gained by calibration and regionalization will depend on the structure and flexiblity of the model in question. Many current models have rigid structures and/or insufficient free parameters and thus cannot be calibrated successfully (Mendoza et al., 2015a). Moreover, for climate projections one should bear in mind that calibrated parameters become less valid when the model is subjected to climatic conditions it has

35 never seen before (Knutti, 2008).

4.5 What is the impact of the forcing data on the simulated runoff?

There are not only strong inter-model differences in the performance patterns but also clear inter-model similarities, suggesting that the forcing data quality imparts a strong limit on the performance. This is most notable for the MAR metric: all models showed negative biases in MAR in snow-dominated regions such as Alaska, the Rocky Mountains, and southern Russia, while they consistently showed positive biases in MAR for the Great Plains (USA) and southern Australia (Figure 1). The high spatial correlation in the performance patterns suggests that these consistent performance patterns may be due to biases in the WFDEI P data, rather than biases in the streamflow observations which are unlikely to be spatially correlated.

It is conceivable that biases are present in the WFDEI P data, because: (i) the monthly CRU dataset, which has been used to correct the WFDEI dataset, is based on only a subset of the available gauges and does not explicitly account for orographic effects; (ii) in sparsely gauged regions the correction using CRU is more likely to deteriorate rather than improve the P estimates; and (iii) the Adam and Lettenmaier (2003) gauge undercatch correction factors are based on interpolation of a very sparse sample of gauges and thus subject to considerable uncertainty. For the conterminous USA we quantified the biases in the WFDEI P data using the high-quality Parameter-elevation Relationships on Independent Slopes Model (PRISM) climatic dataset (Daly et al., 1994), which is based on considerably more gauges than CRU and includes sophisticated corrections for orography. Figure 4a shows the bias in mean annual P from WFDEI relative to that from PRISM, suggesting that the WFDEI P data are indeed subject to large biases. Figure 4b shows the bias in MAR from the MEAN-All ensemble relative to MAR from the observations, revealing a comparable bias pattern, thus confirming that the biases in the WFDEI P propagate in the simulated runoff. The correlation coefficient between the MAR and P bias values is 0.58, indicating a moderately strong relationship. These P biases appear to translate into even more pronounced runoff biases in (semi-)arid regions (notably the northern Great Plains; Figures 4b and 4c) due to the highly non-linear response behavior in these environments (Lidén and Harlin, 2000; Fekete et al., 2004; Van Dijk et al., 2013a). We were unable to quantify the P biases globally since no other independent, global-scale P dataset exists (the WorldClim and CHPclim datasets are likely to exhibit similar biases as the CRU TS3.1 dataset, given that they are based on similar sets of gauges). However, we expect the P biases to be at least similar, if not more severe, outside the well-instrumented conterminous USA (cf. Fekete et al., 2004; Hijmans et al., 2005; Biemans et al., 2009; Zhou et al., 2012; Kauffeldt et al., 2013; Greuell et al., 2015). It should be noted that biases in PET are probably of secondary importance as compared with biases in P (Donohue et al., 2010; Sperna Weiland et al., 2011; Seiller and Anctil, 2015).

The global-scale quantification and reduction of these P biases should be a priority for future research. Satellite-derived P offers unique opportunities in this regard (e.g., Funk et al., 2015) that extend beyond the tropics with the recent launch of the Global Precipitation Measurement (GPM) mission (Smith et al., 2007). Another little-explored way of reducing P uncertainty is by "doing hydrology backwards"; that is, to use information on other hydrological variables—for example, satellite-derived surface soil moisture (e.g., Brocca et al., 2014), streamflow observations (e.g., Adam et al., 2006; Beck et al., 2017), and snow-depth observations (e.g., Cherry et al., 2005)—to reconstruct P through hydrological modeling. Arguably the most important
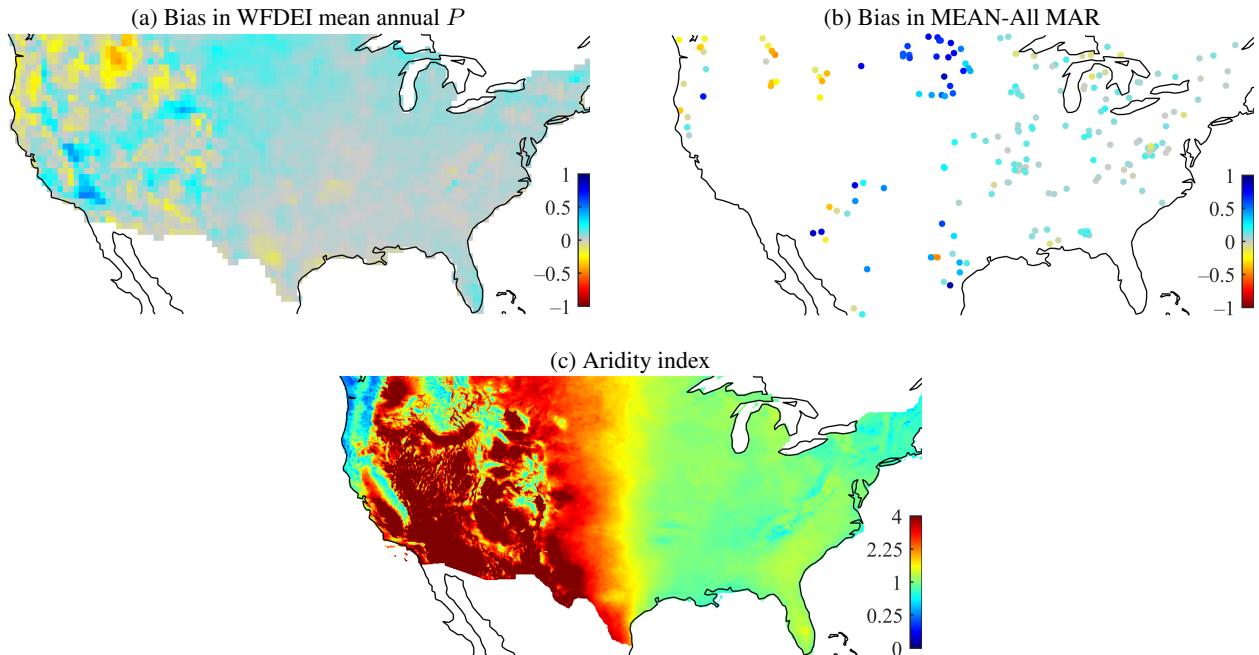
Figure 4. For the conterminous US, (a) the bias in mean annual P from WFDEI relative to PRISM, (b) the bias in MAR from the MEAN-All ensemble relative to the observations, and (c) the aridity index, the ratio of mean annual PET (computed from PRISM air temperature using Hargreaves et al., 1985) to P (PRISM; note the non-linear color scale). Each data point in panel (b) represents a catchment centroid. The bias in (a) and (b) was computed following $B = (X - R)/(X + R)$, where B is the bias, X the uncertain value, and R the reference value. B values range from -1 to 1. A 100 % overestimation results in $B = 1/3$, whereas a 50 % underestimation results in $B = -1/3$.

obstacles to combining multiple data sources are the inconsistent temporal coverage and scale of different data sources and the general lack of error/uncertainty estimates.

Although the models all used the same P data, they used different formulations to compute PET which has likely contributed to differences in simulated runoff among the models in energy-limited regions (Weiß and Menzel, 2008; Kingston et al., 2009; 5   Haddeland et al., 2011; Weedon et al., 2011; Sperna Weiland et al., 2011). However, PET data were available for only four models, which is insufficient to examine whether the PET formulation has had a discernible influence on the simulated runoff, given the numerous other differences in structure and parameterization among the models.

4.6 How valuable are multi-model ensembles?

The multi-model ensemble MEAN-All incorporated all ten models, while MEAN-Best4 incorporated only LISFLOOD, W3RA, 10   WaterGAP3, and HBV-SIMREG (i.e., the four models that performed best in terms of $\overline{OS}$; Table 5). MEAN-All and MEAN-Best4 were found to perform better than all individual models (with the exception of HBV-SIMREG, which has been comprehensively calibrated; Table 5, and Figures 1 and 2). These results highlight the benefits of multi-model ensembles, in line with several previous studies (Ajami et al., 2006; Duan et al., 2007; Viney et al., 2009; Materia et al., 2010; Velázquez et al., 2010;

Gudmundsson et al., 2012a; Xia et al., 2012; Yang et al., 2015). The similar $\overline{\text{OS}}$ scores obtained by MEAN-All and MEAN-Best4 (0.57 and 0.60, respectively; Table 5) suggests that the inclusion of less accurate models has only limited adverse effects. It may be worthwhile for future studies to examine the benefits of more sophisticated multi-model combination techniques involving bias correction or model weighting (e.g., Ajami et al., 2006; Duan et al., 2007; Bohn et al., 2010). These weights can

5 subsequently be transferred from gauged to ungauged areas using regionalization techniques typically used for hydrological model parameters (Blöschl et al., 2013).

HBV-SIMREG differs from the other models because it represents a so-called "multi-parameterization ensemble", which means the model was run multiple (ten) times globally using different (regionalized) parameter sets representing different catchment response behaviors (Beck et al., 2016). HBV-SIMREG obtained slightly better performance than both MEAN-All

10 and MEAN-Best4 overall (Table 5), tentatively suggesting that a multi-parameterization ensemble for a single, sufficiently flexible model provides performance comparable to a multi-model ensemble (cf. Oudin et al., 2006; Yang et al., 2011; Coxon et al., 2014). If this is confirmed, it would negate the need to set up, run, and maintain multiple models, and incentivize the development of a single community hydrological model (cf. Weiler and Beven, 2015) as well as modeling systems allowing selection of alternative model structures (cf. Bierkens, 2015), such as the Framework for Understanding Structural Errors

15 (FUSE; Clark et al., 2008), Noah Multi-Parameterization (Noah-MP; Niu et al., 2011), and SUPERFLEX (Fenicia et al., 2011).

4.7 Do all models show the early bias in runoff timing in snow-dominated catchments previously documented and what is the cause?

With the exception of ORCHIDEE and HBV-SIMREG, all models showed early T50 biases in snow-dominated regions (Sup-

20 plementary material Figure S1.3), indicating that the models produce the spring snowmelt peak early, as has also been reported in several previous studies using different models and forcing data (Lohmann et al., 2004; Slater et al., 2007; Decharme and Douville, 2007; Balsamo et al., 2009; Zaitchik et al., 2010; Beck et al., 2015). The early runoff timing is probably primarily due to P underestimation which leads to insufficient snow accumulation that subsequently melts too quickly (Hancock et al., 2014). The fact that HBV-SIMREG performs well in this regard is probably attributable to the snowfall gauge undercatch

25 correction factor of the model. Indeed, Figure 5 tentatively shows that catchments in which the models strongly underestimate runoff (i.e., negative D_{RC}) generally tend to exhibit an early bias in T50 (i.e., negative D_{T50}) and vice versa. The absence or misrepresentation of certain processes that delay snowmelt runoff in the models may have exacerbated the early runoff timing problem. Examples of such processes include the isothermal phase change of the snowpack, retainment of meltwater in the snowpack in pore spaces, infiltration of meltwater into the soil, meltwater refreezing during cold days and nights, and icejams

30 in rivers. On the whole, more research is needed to ascertain the exact reasons of the early runoff timing.
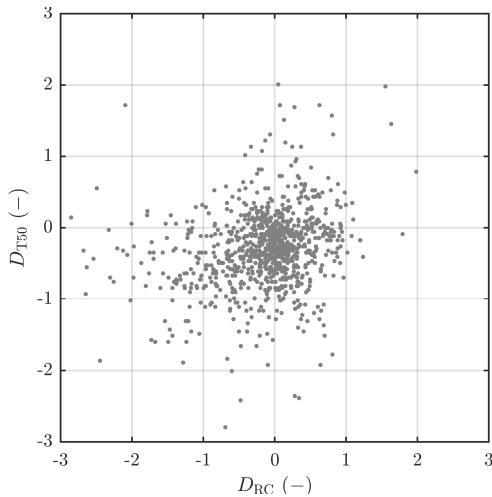
Figure 5. Scatterplot of the difference between simulated (MEAN-All) and observed transformed RC (D_{RC}) versus the difference between simulated (MEAN-All) and observed T50 (D_{T50}) for the catchments ($n = 966$).

5   Conclusions

The runoff estimates from ten state-of-the-art macro-scale hydrological models, all forced with the WFDEI dataset, were evaluated using observations from 966 medium sized catchments around the globe. With reference to the questions posed in the introduction, the following was found:

5   1. The performance differed markedly among models, underscoring the importance of hydrological model uncertainty in addition to climate input uncertainty, and suggesting that model choice should be regarded as a critical step in any hydrological modeling study.

2. The models displayed similar MAR trends, although they were in poor agreement with observed trends. Model-based runoff trends in small-to-medium sized catchments should thus be interpreted with considerable caution.

10   3. Considering only the uncalibrated models, the GHMs performed similarly to the LSMs in rainfall-dominated regions but consistently better than the LSMs in snow-dominated regions, perhaps due to the use of more data-demanding snow routines or the misrepresentation of frozen-soil and snowmelt processes by the LSMs.

4. The models that have been calibrated obtained higher scores for the performance metrics incorporated in the respective objective functions used for calibration.

15   5. The WFDEI P forcing data still appear to contain substantial biases, despite adjustments using gauge observations. These P biases translate into biases in the simulated runoff which are amplified in (semi-)arid regions. In snow-dominated regions there appears to be a consistent underestimation in P and thus simulated runoff.

6. The multi-model ensembles obtained only slightly worse performance than the best (calibrated) model, and the inclusion of less accurate models did not severely degrade the performance. A multi-parameterization ensemble for a single, sufficiently flexible model is easier to realize but we speculate may yield the same performance benefits as a multi-model ensemble.

7. Most models were indeed found to generate the spring snowmelt peak early, probably due to the previously mentioned P underestimation and the absence or misrepresentation of certain processes that delay snowmelt runoff in the models.

Author contributions. H.B. designed and performed the model evaluation and wrote most of the manuscript. A.v.D., A.d.R., E.D., G.F., R.O., and J.S. helped with the interpretation of the results and contributed to writing of the manuscript. H.B., A.v.D., A.d.R., E.D., G.F., and R.O. assisted in running the hydrological models and making available the model output.

Acknowledgements. The Global Runoff Data Centre (GRDC) and the U.S. Geological Survey (USGS) are thanked for providing most of the observed streamflow data. We gratefully acknowledge the modeling groups participating in the eartH2Observe project for providing the simulated runoff data. Lukas Gudmundsson and an anonymous reviewer are thanked for their comments on an earlier draft. This research received funding from the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 603608, "Global Earth Observation for integrated water resource assessment": eartH2Observe. The views expressed herein are those of the authors and do not necessarily reflect those of the European Commission.

References

Adam, J. C. and Lettenmaier, D. P.: Adjustment of global gridded precipitation for systematic bias, Journal of Geophysical Research: Atmospheres, 108, doi:10.1029/2002JD002499, 2003.

Adam, J. C., Clark, E. A., Lettenmaier, D. P., and Wood, E. F.: Correction of global precipitation products for orographic effects, Journal of Climate, 19, 15–38, doi:10.1175/JCLI3604.1, 2006.

Ajami, N. K., Duan, Q., Gao, X., and Sorooshian, S.: Multimodel Combination Techniques for Analysis of Hydrological Simulations: Application to Distributed Model Intercomparison Project Results, Journal of Hydrometeorology, 7, 755–768, 2006.

Andréassian, V., Lerat, J., Loumagne, C., Mathevet, T., Michel, C., Oudin, L., and Perrin, C.: What is really undermining hydrologic science today?, Hydrological Processes, 21, 2819–2822, 2007.

Andréassian, V., Le Moine, N., Perrin, C., Ramos, M. H., Oudin, L., Mathevet, T., Lerat, J., and Berthet, L.: All that glitters is not gold: the case of calibrating hydrological models, Hydrological Processes, 26, 2206–2210, 2012.

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., van den Hurk, B., Hirschi, M., and Betts, A. K.: A revised hydrology for the ECMWF model: verification from field site to terrestrial water storage and impact in the integrated forecast system, Journal of Hydrometeorology, 10, 623–643, 2009.

Balsamo, G., Pappenberger, F., Dutra, E., Viterbo, P., and van den Hurk, B.: A revised land hydrology in the ECMWF model: a step towards daily water flux prediction in a fully-closed water cycle, Hydrological Processes, 25, 1046–1054, 2011.

Bastola, S., Murphy, C., and Sweeney, J.: The role of hydrological modeling uncertainties in climate change impact assessments of Irish river catchments, Advances in Water Resources, 34, 562–576, 2011.

Beck, H. E., van Dijk, A. I. J. M., Miralles, D. G., de Jeu, R. A. M., Bruijnzeel, L. A., McVicar, T. R., and Schellekens, J.: Global patterns in baseflow index and recession based on streamflow observations from 3394 catchments, Water Resources Research, 49, 7843–7863, 2013.

Beck, H. E., van Dijk, A. I. J. M., and de Roo, A.: Global maps of streamflow characteristics based on observations from several thousand catchments, Journal of Hydrometeorology, 16, 1478–1501, 2015.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., and Bruijnzeel, L. A.: Global-scale regionalization of hydrologic model parameters, Water Resources Research, 52, 3599–3622, doi:10.1002/2015WR018247, 2016.

Beck, H. E., van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., and de Roo, A.: MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data, Hydrology and Earth System Sciences, 21, 589–615, 2017.

Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. . L. H., Ménard, C. B., Edwards, J. M., Hendry, M. A., Porson, A., Gedney, N., Mercado, L. M., Sitch, S., Blyth, E., Boucher, O., Cox, P. M., Grimmond, C. S. B., and Harding, R. J.: The Joint UK Land Environment Simulator (JULES), model description — Part 1: Energy and water fluxes, Geoscientific Model Development, 4, 677–699, doi:10.5194/gmd-4-677-2011, 2011.

Beven, K. J.: Changing ideas in hydrology — the case of physically-based models, Journal of Hydrology, 105, 157–172, 1989.

Biemans, H., Hutjes, R. W. A., Kabat, P., Strengers, B. J., Gerten, D., and Rost, S.: Effects of precipitation uncertainty on discharge calculations for main river basins, Journal of Hydrometeorology, 10, 1011–1025, 2009.

Bierkens, M. F. P.: Global hydrology 2015: state, trends, and directions, Water Resources Research, 51, 4923–4947, doi:10.1002/2015WR017173, 2015.

Bierkens, M. F. P., Bell, V. A., Burek, P., Chaney, N., Condon, L. E., David, C. H., de Roo, A., Döll, P., Drost, N., Famiglietti, J. S., Flörke, M., Gochis, D. J., Houser, P., Hut, R., Keune, J., Kollet, S., Maxwell, R. M., Reager, J. T., Samaniego, L., Sudicky, E., Sutanudjaja, E. H., van de Giesen, N., Winsemius, H., and Wood, E.: Hyper-resolution global hydrological modelling: what is next?, Hydrological Processes, 29, 310–320, 2015.

Blöschl, G. and Sivapalan, M.: Scale issues in hydrological modelling: A review, Hydrological Processes, 9, 251–290, 1995.

Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., and Savenije, H., eds.: Runoff Prediction in Ungauged Basins: synthesis across Processes, Places and Scales, Cambridge University Press, New York, US, 2013.

Bock, A. R., Hay, L. E., McCabe, G. J., Markstrom, S. L., and Atkinson, R. D.: Parameter regionalization of a monthly water balance model for the conterminous United States, Hydrology and Earth System Sciences Discussions, 12, 10 023–10 066, 2015.

Bohn, T. J., Sonessa, M. Y., and Lettenmaier, D. P.: Seasonal hydrologic forecasting: do multimodel ensemble averages always yield improvements in forecast skill?, Journal of Hydrometeorology, 11, 1358–1372, 2010.

Bontemps, S., Defourny, P., and van Bogaert, E.: GlobCover 2009, products description and validation report, Tech. rep., ESA GlobCover project, available at: http://ionia1.esrin.esa.int, 2011.

Bosch, J. M. and Hewlett, J. D.: A review of catchment experiments to determine the effect of vegetation changes on water yield and evapotranspiration, Journal of Hydrology, 55, 3–23, 1982.

Breuer, L., Huisman, J. A., Willems, P., Bormann, H., Bronstert, A., Croke, B. F. W., Frede, H., Gräffe, T., Hubrechts, L., Jakeman, A. J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M., and Viney, N. R.: Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use, Advances in Water Resources, 32, 129–146, 2009.

Brocca, L., Ciabatta, L., Massari, C., Moramarco, T., Hahn, S., Hasenauer, S., Kidd, R., Dorigo, W., Wagner, W., and Levizzani, V.: Soil as a natural rain gauge: estimating global rainfall from satellite soil moisture data, Journal of Geophysical Research: Atmospheres, 119, 5128–5141, 2014.

Budyko, M. I.: Climate and life, Academic Press, New York, 1974.

Burek, P., van der Knijff, J., and de Roo, A.: LISFLOOD Distributed Water Balance and Flood Simulation Model Revised User Manual, Tech. Rep. EUR 26162 EN, Joint Research Centre (JRC), Ispra, Italy, doi:10.2788/24719, https://ec.europa.eu/jrc/sites/default/files/lisflood_2013_online.pdf, 2013.

Cherry, J. E., Tremblay, L. B., Déry, S. J., and Stieglitz, M.: Reconstructing solid precipitation from snow depth measurements and a land surface model, Water Resources Research, 41, doi:10.1029/2005WR003965, 2005.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): a modular framework to diagnose differences between hydrological models, Water Resources Research, 44, doi:10.1029/2007WR006735, 2008.

Clark, M. P., Fan, Y., Lawrence, D. M., Adam, J. C., Bolster, D., Gochis, D. J., Hooper, R. P., Kumar, M., Leung, L. R., Mackay, D. S., Maxwell, R. M., Shen, C., Swenson, S. C., and Zeng, X.: Improving the representation of hydrologic processes in Earth System Models, Water Resources Research, 51, 5929–5956, doi:10.1002/2015WR017096, 2015.

Coxon, G., Freer, J., Wagener, T., Odoni, N. A., and Clark, M.: Diagnostic evaluation of multiple hypotheses of hydrological behavior in a limits-of-acceptability framework for 24 UK catchments, Hydrological Processes, 28, 6135–6150, 2014.

Criss, R. E. and Winston, W. E.: Do Nash values have value? Discussion and alternate proposals, Hydrological Processes, 22, 2723–2725, 2008.

Daly, C., Neilson, R. P., and Phillips, D. L.: A statistical-topographic model for mapping climatological precipitation over mountainous terrain, Journal of Applied Meteorology, 33, 140–158, 1994.

Dawdy, D. R. and O'Donnell, T.: Mathematical models of catchment behavior, Journal of the Hydraulics Division, 91, 123–137, 1965.

Debele, B., Srinivasan, R., and Gosain, A. K.: Comparison of Process-Based and Temperature-Index Snowmelt Modeling in SWAT, Water Resources Management, 24, 1065–1088, 2010.

Decharme, B.: Influence of runoff parameterization on continental hydrology: Comparison between the Noah and the ISBA land surface models, Journal of Geophysical Research, 112, D19 108, doi:10.1029/2007JD008463, 2007.

Decharme, B. and Douville, H.: Uncertainties in the GSWP-2 precipitation forcing and their impacts on regional and global hydrological simulations, Climate Dynamics, 27, 695–713, 2006.

Decharme, B. and Douville, H.: Global validation of the ISBA sub-grid hydrology, Climate Dynamics, 29, 21–37, 2007.

Decharme, B., Boone, A., Delire, C., and Noilhan, J.: Local evaluation of the Interaction between Soil Biosphere Atmosphere soil multilayer diffusion scheme using four pedotransfer functions, Journal of Geophysical Research: Atmospheres, 116, D20 126, 2011.

Decharme, B., Martin, E., and Faroux, S.: Reconciling soil thermal and hydrological lower boundary conditions in land surface models, Journal of Geophysical Research: Atmospheres, 118, 7819–7834, 2013.

Dirmeyer, P. A.: A history and review of the Global Soil Wetness Project (GSWP), Journal of Hydrometeorology, 12, 729–749, 2011.

Döll, P. and Fiedler, K.: Global-scale modeling of groundwater recharge, Hydrology and Earth System Sciences, 12, 863–885, 2008.

Döll, P., Kaspar, F., and Lehner, B.: A global hydrological model for deriving water availability indicators: model tuning and validation, Journal of Hydrology, 270, 105–134, 2003.

Döll, P., Douville, H., Güntner, A., Müller Schmied, H., and Wada, Y.: Modelling Freshwater Resources at the global scale: challenges and prospects, Surveys in Geophysics, pp. 1–26, doi:10.1007/s10712-015-9343-1, 2015.

Donohue, R. J., McVicar, T. R., and Roderick, M. L.: Assessing the ability of potential evaporation formulations to capture the dynamics in evaporative demand within a changing climate, Journal of Hydrology, 386, 186–197, 2010.

Duan, Q., Schaake, J., and Koren, V.: *A Priori* estimation of land surface model parameters, in: Land Surface Hydrology, Meteorology, and Climate: Observations and Modeling, edited by Lakshmi, V., Albertson, J., and Schaake, J., no. 3 in Water Science and Application, pp. 77–94, AGU, Washington, D.C., US, 2001.

Duan, Q., Gupta, H. V., Sorooshian, S., Rousseau, A. N., and Turcotte, R.: Calibration of watershed models, vol. Water Science and Application, American Geophysical Union, 2004.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., Gusev, Y. M., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O. N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., and Wood, E. F.: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, Journal of Hydrology, 320, 3–17, 2006.

Duan, Q., Ajami, N. K., Gao, X., and Sorooshian, S.: Multi-model ensemble hydrologic prediction using Bayesian model averaging, Advances in Water Resources, 30, 1371–1386, 2007.

Falcone, J. A., Carlisle, D. M., Wolock, D. M., and Meador, M. R.: GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States, Ecology, 91, 621, 2010.

Fekete, B. M., Vörösmarty, C. J., Roads, J. O., and Willmott, C. J.: Uncertainties in precipitation and their impacts on runoff estimates, Journal of Climate, 17, 294–304, 2004.

Fekete, B. M., Looser, U., Pietroniro, A., and Robarts, R. D.: Rationale for monitoring discharge on the ground, Journal of Hydrometeorology, 13, 1977–1986, 2012.

Fenicia, G., Kavetski, D., and Savenije, H. H. G.: Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, Water Resources Research, 47, doi:10.1029/2010WR010174, 2011.

Ferguson, R. I.: Snowmelt runoff models, Progress in Physical Geography, 23, 205–227, 1999.

Franz, K. J., Hogue, T. S., and Sorooshian, S.: Operational snow modeling: Addressing the challenges of an energy balance model for National Weather Service forecasts, Journal of Hydrology, 360, 48–66, 2008.

Freeze, R. A. and Harlan, R. L.: Blueprint for a physically-based, digitally-simulated hydrologic response model, Journal of Hydrology, 9, 237–258, 1969.

Funk, C., Verdin, A., Michaelsen, J., Peterson, P., Pedreros, D., and Husak, G.: A global satellite assisted precipitation climatology, Earth System Science Data, 7, 275–287, doi:10.5194/essd-7-275-2015, 2015.

Giuntoli, I., Vidal, J., Prudhomme, C., and Hannah, D. M.: Future hydrological extremes: the uncertainty from multiple global climate and global hydrological models, Earth System Dynamics, 6, 267–285, 2015a.

Giuntoli, I., Vilarini, G., Prudhomme, C., Mallakpour, I., and Hannah, D. M.: Evaluation of global impact models' ability to reproduce runoff characteristics over the central United States, Journal of Geophysical Research: Atmospheres, 120, 9138–9159, 2015b.

Gosling, S. N., Taylor, R. G., Arnell, N. W., and Todd, M. C.: A comparative analysis of projected impacts of climate change on river runoff from global and catchment-scale hydrological models, Hydrology and Earth System Sciences, 15, 279–294, doi:10.5194/hess-15-279-2011, 2011.

Greuell, W., Andersson, J. C. M., Donnelly, C., Feyen, L., Gerten, D., Ludwig, F., Pisacane, G., Roudier, P., and Schaphoff, S.: Evaluation of five hydrological models across Europe and their suitability for making projections under climate change, Hydrology and Earth System Sciences Discussions, 12, 10 289–10 330, 2015.

Gudmundsson, L. and Seneviratne, S. I.: Towards observation-based gridded runoff estimates for Europe, Hydrology and Earth System Sciences, 19, 2859–2879, 2015.

Gudmundsson, L., Tallaksen, L. M., Stahl, K., Clark, D. B., Dumont, E., Hagemann, S., Bertrand, N., Gerten, D., Heinke, J., Hanasaki, N., Voss, F., and Koirala, S.: Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe, Journal of Hydrometeorology, 13, 604–620, 2012a.

Gudmundsson, L., Wagener, T., Tallaksen, L. M., and Engeland, K.: Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe, Water Resources Research, 48, doi:10.1029/2011WR010911, 2012b.

Güntner, A.: Improvement of global hydrological models using GRACE data, Surveys in Geophysics, 29, 375–397, 2008.

Guo, Z., Dirmeyer, P. A., Gao, X., and Zhao, M.: Improving the quality of simulated soil moisture with a multi-model ensemble approach, Quarterly Journal of the Royal Meteorological Society, 133, 731–747, 2007.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrological Processes, 22, 3802–3813, 2008.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 370, 80–91, 2009.

Gupta, H. V., Perrin, C., Kumar, R., Blöschl, G., Clark, M., Montanari, A., and Andréassian, V.: Large-sample hydrology: a need to balance depth with breadth, Hydrology and Earth System Sciences, 18, 463–477, 2014.

Gustard, A., Bullock, A., and Dixon, J. M.: Low flow estimation in the United Kingdom, Tech. Rep. 108, Institute of Hydrology, Wallingford, UK, 1992.

Haddeland, I., Clark, D. B., Franssen, W., F, L., Voß, F., Arnell, N. W., Bertrand, N., Best, M., Folwell, S., Gerten, D., Gomes, S., Gosling, S. N., Hagemann, S., Hanasaki, N., Harding, R., Heinke, J., Kabat, P., Koirala, S., Oki, T., Polcher, J., Stacke, T., Viterbo, P., Weedon, G. P., and Yehm, P.: Multimodel Estimate of the Global Terrestrial Water Balance: Setup and First Results, Journal of Hydrometeorology, 12, 869–884, 2011.

Hancock, S., Huntley, B., Ellis, R., and Baxter, R.: Biases in Reanalysis Snowfall Found by Comparing the JULES Land Surface Model to GlobSnow, Journal of Climate, 27, 624–632, 2014.

Hannah, D. M., Demuth, S., Van Lanen, H. A. J., Looser, U., Prudhomme, C., Rees, G., Stahl, K., and Tallaksen, L. M.: Large-scale river flow archives: importance, current status and future needs, Hydrological Processes, 25, 1191–1200, 2011.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., and Townshend, J. R. G.: High-resolution global maps of 21st-century forest cover change, Science, 342, 850–853, 2013.

Hargreaves, G. L., Hargreaves, G. H., and Riley, J. P.: Irrigation water requirements for Senegal River Basin, Journal of Irrigation and Drainage Engineering, 111, 265–275, 1985.

Harris, I., Jones, P. D., Osborn, T. J., and Lister, D. H.: Updated high-resolution grids of monthly climatic observations—the CRU TS3.10 dataset, International Journal of Climatology, 34, 623–642, 2013.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P., and New, M.: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, Journal of Geophysical Research: Atmospheres, 113, doi:10.1029/2008JD010201, 2008.

He, Y., Bárdossy, A., and Zehe, E.: A review of regionalisation for continuous streamflow simulation, Hydrology and Earth System Sciences, 15, 3539–3553, 2011.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A.: Very high resolution interpolated climate surfaces for global land areas, International Journal of Climatology, 25, 1965–1978, 2005.

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P. A., Uhlenbrook, S., Wagener, T., Winsemius, H. C., Woods, R. A., Zehe, E., and Cudennec, C.: A decade of Predictions in Ungauged Basins (PUB)—a review, Hydrological Sciences Journal, 58, 1198–1255, 2013.

Hunger, M. and Döll, P.: Value of river discharge data for global-scale hydrological modeling, Hydrology and Earth System Sciences, 12, 841–861, 2008.

Jain, S. K. and Sudheer, K. P.: Fitting of hydrologic models: a close look at the Nash-Sutcliffe index, Journal of Hydrologic Engineering, 13, 981–986, 2008.

Jiménez, C., Prigent, C., Mueller, B., Seneviratne, S. I., McCabe, M. F., Wood, E. F., Rossow, W. B., Balsamo, G., Betts, A. K., Dirmeyer, P. A., Fisher, J. B., Jung, M., Kanamitsu, M., Reichle, R. H., Reichstein, M., Rodell, M., Sheffield, J., Tu, K., and Wang, K.: Global intercomparison of 12 land surface heat flux estimates, Journal of Geophysical Research, 116, D02 102, doi:10.1029/2010JD014545, 2011.

Kauffeldt, A., Halldin, S., Rodhe, A., Xu, C.-Y., and Westerberg, I. K.: Disinformative data in large-scale hydrological modelling, Hydrology and Earth System Sciences, 17, 2845–2013, 2013.

Kauffeldt, A., Wetterhall, F., Pappenberger, F., Salamon, P., and Thielen, J.: Technical review of large-scale hydrological models for implementation in operational flood forecasting schemes on continental level, Environmental Modelling & Software, 75, 68–76, doi:10.1016/j.envsoft.2015.09.009, 2016.

Kingston, D. G., Todd, M. C., Taylor, R. G., Thompson, J. R., and Arnell, N. W.: Uncertainty in the estimation of potential evapotranspiration under climate change, Geophysical Research Letters, 36, doi:10.1029/2009GL040267, 2009.

Kleidon, A., Renner, M., and Porada, P.: Estimates of the climatological land surface energy and water balance derived from maximum convective power, Hydrology and Earth System Sciences, 18, 2201–2218, 2014.

Klemeš, V.: Operational testing of hydrological simulation models, Hydrological Sciences Journal, 31, 13–24, 1986.

Knutti, R.: Should we believe model predictions of future climate change?, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 366, 4647–4664, 2008.

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S., and Prentice, I. C.: A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, Global Biogeochemical Cycles, 19, doi:10.1029/2003GB002199, 2005.

Lehner, B.: Derivation of watershed boundaries for GRDC gauging stations based on the HydroSHEDS drainage network, Tech. Rep. 41, Global Runoff Data Centre (GRDC), Federal Institute of Hydrology (BfG), Koblenz, Germany, 2012.

Lehner, B., Reidy Liermann, C., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., Magome, J., Nilsson, C., Robertson, J. C., Rödel, R., Sindorf, N., and Wisser, D.: High resolution mapping of the world's reservoirs and dams for sustainable river flow management, Frontiers in Ecology and the Environment, 9, 494–502, 2011.

Lidén, R. and Harlin, J.: Analysis of conceptual rainfall-runoff modelling performance in different climates, Journal of Hydrology, 238, 231–247, 2000.

Linsley, R. K. and Crawford, N. H.: Computation of a synthetic streamflow record on a digital computer, pp. 526–538, International Association of Scientific Hydrology, 1960.

Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., Cosgrove, B. A., Sheffield, J., Duan, Q., Luo, L., Higgins, R. W., Pinker, R. T., and Tarpley, J. D.: Streamflow and water balance intercomparisons of four land surface models in the North American Land Data Assimilation System project, Journal of Geophysical Research: Atmospheres, 109, D07S91, doi:10.1029/2003JD003517, 2004.

Materia, S., Dirmeyer, P. A., Guo, Z., Alessandri, A., and Navarra, A.: The Sensitivity of Simulated River Discharge to Land Surface Representation and Meteorological Forcings, Journal of Hydrometeorology, 11, 334–351, 2010.

McCabe, M. F., Ershadi, A., Jimenez, C., Miralles, D. G., Michel, D., and Wood, E. F.: The GEWEX LandFlux project: evaluation of model evaporation using tower-based and globally-gridded forcing data, Geoscientific Model Development, 9, 283–305, doi:10.5194/gmd-9-283-2016, 2016.

McDonnell, J. J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., Hinz, C., Hooper, R., Kirchner, J., Roderick, M. L., Selker, J., and Weiler, M.: Moving beyond heterogeneity and process complexity: a new vision for watershed hydrology, Water Resources Research, 43, W07 301, doi:10.1029/2006WR005467, 2007.

Mendoza, P. A., Clark, M. P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G., and Gupta, H.: Are we unnecessarily constraining the agility of complex process-based models?, Water Resources Research, 51, 716–728, doi:10.1002/2014WR015820, 2015a.

Mendoza, P. A., Clark, M. P., Mizukami, N., Newman, A. J., Barlage, M., Gutmann, E. D., Rasmussen, R. M., Rajagopalan, B., Brekke, L. D., and Arnold, J. R.: Effects of hydrologic model choice and calibration on the portrayal of climate change impacts, Journal of Hydrometeorology, 16, 762–780, 2015b.

Milly, P. C. D., Dunne, K. A., and Vecchia, A. V.: Global pattern of trends in streamflow and water availability in a changing climate., Nature, 438, 347–350, doi:10.1038/nature04312, 2005.

Minville, M., Cartier, D., Guay, C., Leclaire, L.-A., Audet, C., Le Digabel, S., and Merleau, J.: Improving process representation in conceptual hydrological model calibration using climate simulations, Water Resources Research, 50, 5044–5073, 2014.

Miralles, D. G., Jímenez, C., Jung, M., Michel, D., Ershadi, A., McCabe, M. F., Hirschi, M., Martens, B., Dolman, A. J., Fisher, J. B., Mu, Q., Seneviratne, S. I., Wood, E. F., and Fernaíndez-Prieto, D.: The WACMOS-ET project — Part 2: evaluation of global terrestrial evaporation data sets, Hydrology and Earth System Sciences, 20, 823–842, doi:10.5194/hess-20-823-2016, 2015.

Monk, W. A., Wood, P. J., Hannah, D. M., and Wilson, D. A.: Selection of river flow indices for the assessment of hydroecological change, River Research and Applications, 23, 113–122, 2007.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—a discussion of principles, Journal of Hydrology, 10, 282–290, 1970.

Nasonova, O. N., Gusev, Y. M., and Kovalev, Y. E.: Investigating the Ability of a Land Surface Model to Simulate Streamflow with the Accuracy of Hydrological Models: a Case Study Using MOPEX Materials, Journal of Hydrometeorology, 10, 1128–1150, 2009.

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and Xia, Y.: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements, Journal of Geophysical Research: Atmospheres, 116, doi:10.1029/2010JD015139, 2011.

Ol'dekop, E. M.: Ob isparenii s poverknosti rechnykh basseinov (On evaporation from the surface of river basins), Transactions on Meteorological Observations, University of Tartu 4, 1911.

Olden, J. D. and Poff, N. L.: Redundancy and the choice of hydrologic indices for characterizing streamflow regimes, River Research and Applications, 19, 101–121, 2003.

Orth, R. and Seneviratne, S.: Introduction of a simple-model-based land surface dataset for Europe, Environmental Research Letters, 10, doi:10.1088/1748-9326/10/4/044012, 2015.

Orth, R., Staudinger, M., Seneviratne, S. I., Seibert, J., and Zappa, M.: Does model performance improve with complexity? A case study with three hydrological models, Journal of Hydrology, 523, 147–159, doi:10.1016/j.jhydrol.2015.01.044, 2015.

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., and Michel, C.: Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations, Water Resources Research, 42, doi:10.1029/2005WR004636, 2006.

Parajka, J., Viglione, A., Rogger, M., Salinas, J. L., Sivapalan, M., and Blöschl, G.: Comparative assessment of predictions in ungauged basins — Part 1: Runoff-hydrograph studies, Hydrology and Earth System Sciences, 17, 1783–1795, 2013.

Peel, M. C., Chiew, F. H. S., Western, A. W., and McMahon, T. A.: Extension of unimpaired monthly streamflow data and regionalisation of parameter values to estimate streamflow in ungauged catchments, report prepared for the Australian National Land and Water Resources Audit. Centre for Environmental Applied Hydrology, University of Melbourne, Australia, 2000.

Pike, J. G.: The estimation of annual run-off from meteorological data in a tropical climate, Journal of Hydrology, 2, 116–123, 1964.

Pilgrim, D. H., Chapman, T. G., and Doran, D. G.: Problems of rainfall-runoff modelling in arid and semiarid regions, Hydrological Sciences Journal, 33, 379–400, 1988.

Porporato, A., Daly, E., and Rodriguez-Iturbe, I.: Soil water balance and ecosystem response to climate change, The American Naturalist, 164, 625–632, 2004.

Prudhomme, C., Parry, S., Hannaford, J., Clark, D. B., Hagemann, S., and Voss, F.: How Well Do Large-Scale Models Reproduce Regional Hydrological Extremes in Europe?, Journal of Hydrometeorology, 12, 1181–1204, 2011.

Prudhomme, C., Giuntoli, I., Robinson, E. L., Clark, D. B., Arnell, N. W., Dankers, R., Fekete, B. M., Franssen, W., Gerten, D., Gosling, S. N., Hagemann, S., Hannah, D. M., Kim, H., Masaki, Y., Satoh, Y., Stacke, T., Wada, Y., and Wisser, D.: Hydrological droughts in the 21st century, hotspots and uncertainties from a global multimodel ensemble experiment, Proceedings of the National Academy of Sciences of the United States of America, 111, 3262–3267, 2014.

Razavi, T. and Coulibaly, P.: Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods, Journal of Hydrologic Engineering, 18, 958–975, 2013.

Rockwood, D. M.: Streamflow synthesis and reservoir regulation, Engineering Studies Project 171 Technical Bulletin No. 22, U.S. Army Engineer Division, North Pacific, Portland, Oregon, 1964.

Rosbjerg, D. and Madsen, H.: Concepts of Hydrologic Modeling, in: Encyclopedia of Hydrological Sciences, chap. 10, John Wiley & Sons, doi:10.1002/047048944, 2006.

Rosero, E., Gulden, L. E., and Yang, Z.: Ensemble Evaluation of Hydrologically Enhanced Noah-LSM: partitioning of the Water Balance in High-Resolution Simulations over the Little Washita River Experimental Watershed, Journal of Hydrometeorology, 12, 45–64, 2011.

Schaefli, B. and Gupta, H. V.: Do Nash values have value?, Hydrological Processes, 21, 2075–2080, 2007.

Schellekens, J., Dutra, E., Martínez-de la Torre, A., van Dijk, G. B. A., Sperna Weiland, F., Minvielle, M., Calvet, J.-C., Decharme, B., Eisner, S., Fink, G., Flörke, M., Peßenteiner, S., van Beek, R., Polcher, J., Beck, H., Orth, R., Calton, B., ad W. Dorigo, S. B., and Weedon, G. P.: A global water resources ensemble of hydrological models: the eartH2Observe Tier-1 dataset, Earth System Science Data Discussions, pp. 1–35, doi:10.5194/essd-2016-55, 2016.

Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N. W., Clark, D. B., Dankers, R., Eisner, S., Fekete, B. M., Colón-González, F. J., Gosling, S. N., Kim, H., Liu, X., Masaki, Y., Portmann, F. T., Satoh, Y., Stacke, T., Tang, Q., Wada, Y., Wisser, D., Albrecht, T., Frieler, K., Piontek, F., Warszawski, L., , and Kabat, P.: Multimodel assessment of water scarcity under climate change, Proceedings of the National Academy of Sciences of the United States of America, 111, 3245–3250, 2013.

Schlosser, C. A. and Gao, X.: Assessing Evapotranspiration Estimates from the Second Global Soil Wetness Project (GSWP-2) Simulations, Journal of Hydrometeorology, 11, 880–897, 2010.

Seiller, G. and Anctil, F.: How do potential evapotranspiration formulas influence hydrological projections?, Hydrological Sciences Journal, advance online publication, doi:10.1080/02626667.2015.1100302, 2015.

Sen, P. K.: Estimates of the regression coefficient based on Kendall's tau, Journal of the American Statistical Association, 63, 1379–1389, 1968.

Shafii, M. and Tolson, B. A.: Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives, Water Resources Research, 51, 3796–3814, doi:10.1002/2014WR016520, 2015.

Siebert, S., Döll, P., Hoogeveen, J., Faures, J., Frenken, K., and Feick, S.: Development and validation of the global map of irrigation areas, Hydrology and Earth System Sciences, 9, 535–547, doi:10.5194/hess-9-535-2005, 2005.

Singh, V. P., ed.: Computer models of watershed hydrology, Water Resources Publications, Colorado, USA, 1995.

Singh, V. P. and Frevert, D. K., eds.: Mathematical models of large watershed hydrology, Water Resources Publications, Colorado, USA, 2002.

Sivapalan, M.: Prediction in ungauged basins: a grand challenge for theoretical hydrology, Hydrological Processes, 17, 3163–3170, 2003.

Slater, A. G., Schlosser, C. A., Desborough, C. E., Pitman, A. J., Henderson-Sellers, A., Robock, A., Vinnikov, K. Y., Entin, J., Mitchell, K., Chen, F., Boone, A., Etchevers, P., Habets, F., Noilhan, J., Braden, H., Cox, P. M., de Rosnay, P., Dickinson, R. E., Yang, Z., Dai, Y., Zeng, Q., Duan, Q., Koren, V., Schaake, S., Gedney, N., Gusev, Y. M., Nasonova, O. N., Kim, J., Kowalczyk, E. A., Shmakin, A. B., Smirnova, T. G., Verseghy, D., Wetzel, P., , and Xue, Y.: The representation of snow in land surface schemes: results from PILPS 2(d), Journal of Hydrometeorology, 2, 7–25, 2001.

Slater, A. G., Bohn, T. J., McCreight, J. L., Serreze, M. C., and Lettenmaier, D. P.: A multimodel simulation of pan-Arctic hydrology, Journal of Geophysical Research: Biogeosciences, 112, doi:10.1029/2006JG000303, 2007.

Smith, E. A., Asrar, G. R., Furuhama, Y., Ginati, G., Kummerow, C., Levizzani, V., Mugnai, A., Nakamura, K., Adler, R., Casse, V., Cleave, M., Debois, M., John, J., Entin, J., Houser, P., Iguchi, T., Kakar, R., Kaye, J., Kojima, M., Lettenmaier, D., Luther, M., Mehta, A., Morel, P., Nakazawa, T., Neeck, S., Okamoto, K., Oki, R., Raju, G., Shepherd, M., Stocker, E., Testud, J., and Wood, E.: The International Global Precipitation Measurement (GPM) program and mission: An overview, in: Measuring Precipitation From Space, pp. 611–653, Springer, New York, 2007.

Sooda, A. and Smakhtin, V.: Global hydrological models: a review, Hydrological Sciences Journal, 60, doi:10.1016/j.jhydrol.2012.09.002, 2015.

Sperna Weiland, F. C., Tisseuil, C., Dürr, H. H., Vrac, M., and van Beek, L. P. H.: Selecting the optimal method to calculate daily global reference potential evaporation from CFSR reanalysis data, Hydrology and Earth System Sciences, 16, 983–1000, 2011.

Stahl, K., Tallaksen, L. M., Gudmundsson, L., and Christensen, J. H.: Streamflow Data from Small Basins: A Challenging Test to High-Resolution Regional Climate Modeling, Journal of Hydrometeorology, 12, 900–912, 2011.

Stahl, K., Tallaksen, L. M., Hannaford, J., and van Lanen, H. A. J.: Filling the white space on maps of European runoff trends: estimates from a multi-model ensemble, Hydrology and Earth System Sciences, 16, 2035–2047, 2012.

Stewart, I. T., Cayan, D. R., and Dettinger, M. D.: Changes toward Earlier Streamflow Timing across Western North America, Journal of Climate, 18, 1136–1155, 2005.

Sugawara, M.: The flood forecasting by a series storage type model, in: Int. Symposium Floods and their Computation, International Association of Hydrologic Sciences, 1967.

Tait, A., Henderson, R., Turner, R., and Zheng, X.: Thin plate smoothing spline interpolation of daily rainfall for New Zealand using a climatological rainfall surface, International Journal of Climatology, 26, 2097–2115, 2006.

Tebaldi, C. and Knutti, R.: The use of the multi-model ensemble in probabilistic climate projections, Philosophical Transactions of The Royal Society of London, Series A, 365, 2053–2075, 2007.

Teutschbein, C. and Seibert, J.: Regional Climate Models for Hydrological Impact Studies at the Catchment Scale: A Review of Recent Modeling Strategies, Geography Compass, 4, 834–860, 2010.

Trambauer, P., Maskeya, S., Winsemius, H., Werner, M., and Uhlenbrook, S.: A review of continental scale hydrological models and their suitability for drought forecasting in (sub-Saharan) Africa, Physics and Chemistry of the Earth, 66, 16–26, 2013.

Van Beek, L. P. H. and Bierkens, M. F. P.: The Global Hydrological Model PCR-GLOBWB: conceptualization, Parameterization and Verification, Tech. rep., Utrecht University, http://vanbeek.geo.uu.nl/suppinfo/vanbeekbierkens2009.pdf, 2009.

Van Dijk, A. I. J. M.: AWRA Technical Report 3, Landscape Model (version 0.5) Technical Description, Tech. rep., WIRADA/CSIRO Water for a Healthy Country Flagship, Canberra, Australia, http://www.clw.csiro.au/publications/waterforahealthycountry/2010/wfhc-aus-water-resources-assessment-system.pdf, 2010.

Van Dijk, A. I. J. M., Beck, H. E., Crosbie, R. S., de Jeu, R. A. M., Liu, Y. Y., Podger, G. M., Timbal, B., and Viney, N. R.: The Millennium Drought in southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society, Water Resources Research, 49, 1040–1057, 2013a.

Van Dijk, A. I. J. M., Peña-Arancibia, J. L., Wood, E. F., Sheffield, J., and Beck, H. E.: Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide, Water Resources Research, 49, 2729–2746, 2013b.

Velázquez, J. A., Anctil, F., and Perrin, C.: Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments, Hydrology and Earth System Sciences, 14, 2303–2317, doi:10.5194/hess-14-2303-2010, 2010.

Verzano, K.: Climate change impacts on flood related hydrological processes: Further development and application of a global scale hydrological model, Tech. rep., Max Planck Institute for Meteorology, Hamburg, Germany, 2009.

Viney, N. R., Bormann, H., Breuer, L., Bronstert, A., Croke, B. F. W., Frede, H., Gräffe, T., Hubrechts, L., Jakeman, A. J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M., and Willems, P.: Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) II: Ensemble combinations and predictions, Advances in Water Resources, 32, 147–158, 2009.

Vis, M., Knight, R., Pool, S., Wolfe, W., and Seibert, J.: Model calibration criteria for estimating ecological flow characteristics, Water, 7, 2358–2381, 2015.

Wagener, T.: Evaluation of catchment models, Hydrological Processes, 17, 3375–3378, 2003.

Wandishin, M. S., Mullen, S. L., Stensrud, D. J., and Brooks, H. E.: Evaluation of a Short-Range Multimodel Ensemble System, Monthly Weather Review, 129, 729, 2001.

Weedon, G. P., Gomes, S., Viterbo, P., Shuttleworth, W. J., Blyth, E., Österle, H., Adam, J. C., Bellouin, N., Boucher, O., and Best, M.: Creation of the WATCH Forcing Data and Its Use to Assess Global and Regional Reference Crop Evaporation over Land during the Twentieth Century, Journal of Hydrometeorology, 12, 823–848, 2011.

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., and Viterbo, P.: The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, Water Resources Research, 50, 7505–7514, 2014.

Weiler, M. and Beven, K.: Do we need a community hydrological model?, Water Resources Research, 51, 7777–7784, doi:10.1002/2014WR016731, 2015.

Weiß, M. and Menzel, L.: A global comparison of four potential evapotranspiration equations and their relevance to stream flow modelling in semi-arid environments, Advances in Geosciences, 18, 15–23, doi:10.5194/adgeo-18-15-2008, 2008.

Westerberg, I. K. and McMillan, H. K.: Uncertainty in hydrological signatures, Hydrology and Earth System Sciences, 19, 3951–3968, 2015.

WMO: Intercomparison of conceptual models used in operational hydrological forecasting, Tech. Rep. WMO no. 429, Operational Hydrology Report no. 7, World Meteorological Organization, Geneva, Switzerland, 1975.

WMO: Results of an intercomparison of models of snowmelt runoff, Tech. Rep. WMO no. 646, Operational Hydrology Report no. 23, World Meteorological Organization, Geneva, Switzerland, 1986.

WMO: Simulated real-time intercomparison of hydrological models, Tech. Rep. WMO no. 779, Operational Hydrology Report no. 38, World Meteorological Organization, Geneva, Switzerland, 1992.

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., Alonge, C., H, W., Meng, J., Livneh, B., Duan, Q., and Lohmann, D.: Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project

phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow, Journal of Geophysical Research: Atmospheres, 117, D03 110, doi:10.1029/2011JD016048, 2012.

Xia, Y., Sheffield, J., Ek, M. B., Dong, J., Chaney, N., Wei, H., and Wood, J. M. E. F.: Evaluation of multi-model simulated soil moisture in NLDAS-2, Journal of Hydrology, 512, 107–125, doi:10.1016/j.jhydrol.2014.02.027, 2014.

5 Yang, H., Piao, S., Zeng, Z., Ciais, P., Yin, Y., Friedlingstein, P., Sitch, S., Ahlström, A., Guimberteau, M., Huntingford, C., Levis, S., Levy, P. E., Huang, M., Li, Y., Li, X., Lomas, M. R., Peylin, P., Poulter, B., Viovy, N., Zaehle, S., Zeng, N., Zhao, F., and Wang, L.: Multicriteria evaluation of discharge simulation in dynamic global vegetation models, Journal of Geophysical Research: Atmospheres, 120, 7488–7505, 2015.

Yang, Z., Niu, G., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Longuevergne, L., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and
10 Xia, Y.: The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins, Journal of Geophysical Research: Atmospheres, 116, doi:10.1029/2010JD015140, 2011.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resources Research, 44, doi:10.1029/2007WR006716, 2008.

Zaitchik, B. F., Rodell, M., and Olivera, F.: Evaluation of the Global Land Data Assimilation System using global river discharge data and a
15 source-to-sink routing scheme, Water Resources Research, 46, doi:10.1029/2009WR007811, 2010.

Zeinivand, H. and De Smedt, F.: Hydrological Modeling of Snow Accumulation and Melting on River Basin Scale, Water Resources Management, 23, 2271–2287, 2009.

Zhang, L., Dawes, W. R., and Walker, G. R.: Response of mean annual evapotranspiration to vegetation changes at catchment scale, Water Resources Research, 37, 701–708, doi:10.1029/2000WR900325, 2001.

20 Zhang, Y., Zheng, H., Chiew, F., Peña-Arancibia, J., and Zhou, X.: Evaluating regional and global hydrological models against streamflow and evapotranspiration measurements, Journal of Hydrometeorology, early online publication, doi:10.1175/JHM-D-15-0107.1, 2016.

Zhou, X., Zhang, Y., Wang, Y., Zhang, H., Vaze, J., Zhang, L., Yang, Y., and Zhou, Y.: Benchmarking global land surface models against the observed mean annual runoff from 150 large basins, Journal of Hydrology, 470–471, 269–279, 2012.